WARWICK
ECONOMICS

Did railways affect literacy? Evidence from India

Latika Chaudhary & James Fenske

(This paper also appears as CAGE Discussion paper 529)

December 2020 No: 1320

Warwick Economics Research Papers

ISSN 2059-4283 (online)
ISSN 0083-7350 (print)

DID RAILWAYS AFFECT LITERACY? EVIDENCE FROM INDIA

LATIKA CHAUDHARY* AND JAMES FENSKE†

Abstract. We study the effect of railroads, the single largest public investment in colonial India, on human capital. Using district-level data on literacy, we find railroads had positive effects on literacy, in particular on male and English literacy. We employ two identification strategies. First, we exploit synthetic panel variation contained in cohort-specific literacy rates due to differences in the timing of railroad exposure of different cohorts within the same district and census year. We find a one standard deviation increase in railroad exposure raises literacy by 0.29 standard deviations. Second, we use distance from an early railway plan as an instrument for district railway exposure in the cross section and find results of similar magnitude. We show that railroads increased literacy by raising secondary, rather than primary, schooling. Our mediation analysis suggests that non-agricultural income and opportunities for skilled employment are important mechanisms, while agricultural income is not.

Keywords: Colonialism, Railways, Literacy.
JEL Codes: N75, N35, R40

*Associate Professor, Graduate School of Defense Management, Naval Postgraduate School
†Professor, Department of Economics, University of Warwick
E-mail addresses: lhartman@nps.edu, J.Fenske@warwick.ac.uk.
Date: December 21, 2020.
We are grateful to an audience at the Economic History Association Meetings for comments, to Jinlin Wei for research assistance, and to the Centre for Competitive Advantage in the Global Economy and the Leverhulme Trust for funding.

1. Introduction

By 1900, the rail network in colonial India was the fourth largest in the world, covering more than 40,000 kilometers across more than 200 districts (Bogart and Chaudhary, 2016). In striking contrast, public education was poorly funded and saw marginal progress under British rule. Education was an insignificant line item in the government budget, a mere 1.7% compared to 21% for railroads in 1881 (Government of India, 1887). Education levels were low; in 1891 only 9.6% of primary school-age children were in school (Chaudhary, 2016). Scholars and policymakers continue to debate the relative importance of demand and supply factors in explaining levels of schooling, both in colonial India and in the present (Great Britain, 1929; Banerjee and Duflo, 2011). According to official opinion, demand for basic education was low in rural colonial India, where children helped parents in the field (Chaudhary, 2016). By increasing trade, agricultural income, and other labor market opportunities, railways in India may have increased demand for schooling, even in the absence of supply-side, government-led interventions. In this paper, we ask whether there was a demand-driven increase in education in colonial India in response to the extension of the railway network.

Using decennial census data on Indian literacy from 1881 to 1921, we estimate the effect of railroads on total, male, female, and English literacy at the district level. Railroad construction began in the 1850s, and 52% of British Indian districts were connected to a railroad by 1881. This increased to 87% in 1901 and then 96% in 1921. From 1881, data on literacy became available in the decennial censuses. The early censuses (1881 to 1901) cannot be compared to each other, or to later censuses, due to different enumeration standards and to changing age bins used to define cohorts. We therefore use two principal strategies to identify the effect of railroads. The first exploits panel-like variation across birth cohorts within a given district in a given census year. The second exploits cross-sectional variation across districts in a given census year.

In our first approach, we estimate the differential effect of exposure to railroads across cohorts within districts using the censuses of 1911 and 1921, years with comparable literacy data by cohort. Since 93% of districts are connected to the railroad by 1911, we construct upper and lower bound estimates of the cumulative number of years a cohort is exposed to railroads. Our first measure is the cumulative number of years a railroad was present in a district before the youngest member of the cohort of interest reached age 6. Our second is the cumulative number of years a railroad was present in a district before the youngest member of the cohort of interest reached age 12. These two ages capture the regular start and end of primary school. We take these as alternative measures of railroad exposure. We create a synthetic panel by including district fixed effects that absorb any unobservable, time-invariant factors that correlate with the timing of connection to railroads and that may also correlate with differences in literacy across districts. We also include cohort × province and census year × province fixed effects. These control for both national and provincial factors that affect cohort literacy flexibly over time. These fixed effects mitigate common endogeneity concerns relating to railroad exposure.

In our second approach, we use an instrumental variables (IV) strategy, which exploits cross-sectional variation in the years of railroad exposure in each census between 1881 and 1921. Building on recent techniques in the transportation literature (Reading and Turner 2014), we construct our instrument using a 1852 plan proposed by Major Kennedy, consulting engineer to the Government of India (GOI), which predated railway construction (Davidson, 1868). Kennedy proposed low-cost routes favouring gentle terrain over more direct routes. Our instrument measures the distance between a district and this plan. Distance from the Kennedy plan strongly predicts when a district received a railroad and hence years of exposure to the railroad. Our exclusion restriction assumes distance to the Kennedy plan only affects literacy via railroads and is uncorrelated with unobserved determinants of literacy once we control for observable differences in geography, crop suitability, and religion across districts. In our cross-sectional analyses, we complement our IV strategy with alternatives, including ordinary least squares (OLS), alternative instruments based on spanning trees that connect cities that existed prior to railroad construction and military cantonments that existed early in the construction period, fixed effects for grid cells, and matching estimates.

We find positive and significant effects of railroads on literacy, in particular male and English literacy in the synthetic panel regressions. A standard deviation increase in railroad exposure (17 years) increases total literacy by 0.29 standard deviations for total, 0.31 for male, and 0.25 for male English literacy. We find small and insignificant effects on female literacy. We find similar effects using the second measure of exposure at 0.26 standard deviations for total and 0.29 standard deviations for male literacy. In our cross-sectional regressions, we find positive and significant effects of railroad exposure. Standardised coefficients suggest effect sizes ranging from 0.15 to 0.37 standard deviations depending on the measure of literacy, the census year used, and the specific statistical model. These results support the synthetic panel findings, with the exception of a positive impact on female literacy.

Why did railroads increase education? We complement our literacy results using a novel panel data set we have created on primary and secondary enrolment at the district level. These data cover around 179 British Indian districts in 1901 and 1911.[1] We use both our panel and cross-sectional approaches. In these data, a one standard deviation increase in railroad exposure increases secondary enrolment by 0.45 standard deviations, with larger standardised effects in panel models compared to the cross-sectional regressions. We find

[1]We also have data on a smaller set of districts for 1894, 1897 and 1905.

small and insignificant effects on primary enrolment. Our results are, then, driven by an increase in secondary schooling, not primary schooling.

What links railroads with greater secondary schooling and literacy? Past work has shown that railroads increased agricultural income (Donaldson 2018), which in theory can increase literacy if schooling is a normal good. But, employing mediation approaches from Dippel et al. (2020) and from Imai et al (2010a; 2010b), we find agricultural income is not a significant mediator. Rather, our analysis shows that income taxes and service sector employment are key mediators through which railroads increase literacy and enrolment. These measures capture non-agricultural income and the returns to skill. Because we cannot disaggregate the possible mediating effects of rising non-agricultural income, the relaxation of income constraints for families, and increasing returns to literacy, we view these results as suggestive evidence that railroads increased the demand for education via non-agricultural channels.

Colonial railroads had positive and significant effects on human capital, in particular upper-tail human capital such as English literacy and secondary enrolment. Are these effects large? Our effect sizes on railroads are modest compared to comparable estimates from the nineteenth century United States (Atack, Margo and Perlman, 2012). The effects are also modest if we compare them to the impacts of colonial supply-side investments in education. We make a back of the envelope calculation that suggests the cost of making one additional person literate via railroad investment would be more than 500 times as expensive as using public education funding to achieve the same goal. While railroads had many benefits other than increasing education, they could not make a significant dent in Indian illiteracy given the effect sizes we estimate.

We conduct several robustness checks. First, in the synthetic panel estimation, we drop the cohort aged 20 and above because our approach will lead to the most mis-measurement of railway exposure for the older ages in this cohort. Second, we exploit variation across cohorts in the 1901 census. There are more districts unconnected to the railroad in 1901 compared to later. By comparing across cohorts in 1901, measurement error in literacy is less of a concern. Third, we estimate spatially adjusted Conley standard errors for our pseudo-panel and cross-sectional regressions. Fourth, we test for outliers by dropping the four main cities of Bombay, Calcutta, Delhi and Madras that were among the first to be connected. In another test, we drop one province at a time. We find similar results.

1.1. **Contribution.** Our paper contributes to three literatures. First, it contributes to the literature on the effects of transportation infrastructure. One strand of recent work, drawing on South Asia, exploits the expansion of road and highway networks in India with mixed findings (Adukia et al., 2020; Aggarwal, 2018). Adukia et al. (2020) find positive effects of roads on middle school enrolment with null effects on primary school enrolment. Studying a different Indian road construction program, Aggarwal (2018) finds positive enrolment effects on children aged 5 to 14 and negative effects on older children aged 14 to 20. More broadly, this literature has found impacts of transportation infrastructure on outcomes such as economic growth (Banerjee et al., 2020), industrialization (Faber, 2014), urbanization (Baum-Snow et al., 2017; Pérez, 2018), and trade (Morten and Oliveira, 2016), including in historical contexts (Combes et al. 2020; Hanlon and Heblich, 2020). Even though our data come from a colonial, historical context, we find, as in other work, that the effects of colonial railroads are driven more by increasing returns to education and higher non-agricultural income rather than by income or substitution effects of rising agricultural income.

Another strand of this literature focuses on the effects of railroads on price convergence, income, and development. For example, building on classic work by Fogel (1964), Donaldson and Hornbeck (2016) find large effects of railroads on market integration and income. Donaldson (2018) finds colonial Indian railroads reduced transport costs and increased agricultural income, which in turn mitigated the effects of famine (Burgess and Donaldson, 2017). Much of this work focuses on trade and market integration. In the case of US railroads, scholars have looked at the effects on urbanisation, banking, and schooling, among other outcomes (Atack et al., 2010; 2011; 2014). Although there exist papers that look at how current infrastructure expansions affect human capital outcomes such as schooling via labor markets and income, such work is uncommon in the context of railroads. A notable exception is Atack, Margo and Perlman (2012), who study US school attendance in the nineteenth century. Tang (2017), similarly, looks at mortality effects of railroads in Meiji Japan. Our paper looks at the effects of historical railroads on literacy and enrolment, outcomes more commonly examined in work on recent transportation projects (roads and highways) rather than older projects (railroads). We show that the impacts of transportation infrastructure on human capital have not been limited to modern economies.

Second, our paper contributes to the literature on the effects of demand and supply in explaining schooling (Glewwe and Muralidharan, 2016). Many papers estimate the effect of labor demand shifts on education in India with positive effects due to outsourcing facilities (for e.g., Jensen, 2012) and negative effects related to NREGA (Li and Sekhri, 2019). These relate to larger debates on the relative efficacy of demand versus supply interventions in schooling (Banerjee and Dulfo, 2011). On one side, scholars argue that increasing demand for education is sufficient to increase schooling, while other scholars argue public investments are necessary to increase mass education in developing countries. Our paper shows that one of the biggest infrastructure expansions, railroads, had positive effects on literacy and enrolment in India. Yet, these effects are modest and hence not cost-effective if we consider them against increased public funding of education.

Third, our paper relates to the rich literature on Indian railroads. Much has been written about the effects of colonial railroads on trade, with studies showing large effects on price convergence and income, and ambiguous effects on cropping patterns.[2] Unlike prices, wages did not converge across districts due to railroads (Collins, 1999). Railways had positive but small effects on city growth (Fenske, Kala and Wei, 2020). Indian railroads have also featured heavily in nationalist debates on colonisation. Critics argue that the financing of Indian railways delivered excessive returns to British investors, that the network benefitted colonial interests by emphasising port to interior connections over interior to interior connections, and that this stifled economic development in the long run (Dubey, 1965; Satya, 2008). In this view, railways did not industrialise India because they were built to benefit the Empire. An alternative view argues that, although railroads helped colonial interests, they had positive effects on income and returns to British investors were not excessive (Hurd, 1983; Bogart and Chaudhary, 2019). We add to this literature, showing that there were positive effects on schooling, though these favoured men, English literacy, and secondary enrolment.

The rest of the paper is organised as follows: Section 2 provides historical background on literacy and railways in colonial India. Section 3 describes our data, and is followed by an outline of our estimation strategies in Section 4. We report the results in Section 5, discussion potential mechanisms in Section 6 and conclude in Section 6.

2. Historical Background

2.1. **Literacy in colonial India.** As British rule spread in India, former indigenous schools were slowly replaced with a new system of public and private schools (Nurullah and Naik, 1951). Did this transition increase literacy? We have no way of knowing for certain because there are few estimates of literacy before the late 19$^{\text{th}}$ century. Some missionary accounts suggest village schools were common in Bengal in the early 1800s, but offer few specifics on literacy. The British began collecting data on education after the Crown took control in 1858. Indeed, literacy became a standard question on the census beginning with the first census of 1872. Using official reports, Chaudhary (2016) shows that literacy was low but increasing between 1881 and 1941. Men were at least six times as likely to be literate as women. Upper castes were significantly more likely to be literate than lower castes. There were also major differences by religion and large spatial variation.

For individuals aged 10 and over, male literacy in British India increased from 13% in 1901 to 18% in 1931.[3] Female literacy increased from just under 1% to 3% over the same period. The coastal provinces of Bengal, Bombay and Madras had higher literacy in each decade, with male literacy averaging 20% in 1931 compared to 11% in the interior for Central

[2]See Andrabi and Kuehlwein (2010), Hurd (1975), McAlpin (1974), Mukherjee (1980), Studer (2008), and Donaldson (2018).

[3]This discussion draws on Chaudhary (2016).

Provinces and United Provinces. Literacy among Brahmans, the priestly upper caste of Hindus, averaged 33% in 1931, higher than Christians (28%) another group with relatively high literacy. In comparison, literacy was low among marginalised groups, at under 1% for tribal groups known as Scheduled Tribes in modern India. Regional differences were remarkable, with Brahman literacy at 54% in Madras compared to 16% in the United Provinces.

What accounts for the disappointing performance of literacy? Colonial officials pointed to low demand for education in a rural country like India, high illiteracy among parents, and cultural taboos against Hindu and Muslim girls attending school (Nurullah and Naik, 1951). While demand for basic literacy may be lower in agricultural societies than in industrial countries, it is hard to argue that demand for basic literacy was lower in India than countries at similar levels of development. For example, Lindert (2004) and Chaudhary (2016) show that primary school enrolment in colonial India between 1860 and 1920 was far below rates in Brazil, Mexico, Japan and Chile, a reasonable comparison set based on GDP per capita.

A more plausible first-order explanation was the poor funding of public schooling. A large share of public funding for rural primary schools came from surcharges on land revenues. But these surcharges were fixed in nominal terms, and land revenues did not keep pace with the increase in agricultural incomes. Moreover, education was a small share of the government budget, increasing from 1.5% in 1882 to 7.7% in 1932. Public education spending accounted for less than 1% of national income as late as 1931. Public expenditures on education under the Raj were lower than in other developing countries and in other British colonies (Chaudhary 2009). Against this backdrop of low but varying literacy, few scholars have looked at the effects of demand shifters in explaining levels of schooling. Our paper is a step in this direction.

2.2. **Railroads in colonial India.** Unlike schooling, the British were early promoters of railroads in India, building an extensive rail network.[4] The first passenger line opened in 1853, connecting Bombay to Thane, a distance of 20 miles. Prompted by mercantile interests in Britain, the early lines connected the ports of Bombay, Calcutta and Madras to the interior. Given the few good roads and navigable rivers, British firms hoped railways would lower the costs of exporting raw cotton from India and of importing British manufactured goods to new Indian markets (Thorner 1951; 1955). Indeed, the British believed goods traffic would significantly exceed any passenger traffic. They proved to be wrong, with passenger traffic accounting for 60% of revenues.

Indian railways were built by British firms with British financing, albeit subsidised by a guaranteed dividend backed by the GOI. Such firms were the main players up to the 1870s, when the GOI began to build lines. This was followed by mixed public-private partnerships

[4]There is a large literature on Indian railways. Edited volumes by Kerr (2001, 2007) offer an excellent introduction to the main issues, while Sanyal (1930) offers a detailed history of railway development.

in the 1880s. These partnerships were the norm until the 1920s (Sanyal 1930). Route mileage expanded quickly in the early decades, especially from 1881 to 1901. Total route miles increased from 9,893 in 1881 to 17,283 to 1891, 25,365 in 1901, 32,839 in 1911 and then 37,266 in 1921 (Bogart and Chaudhary, 2016).

Figure 1 maps the spread of the network from 1881 to 1921. The important ports were connected to the interior before 1881. Many lines crossed the densely populated Indo-Gangetic plain with fewer interior lines in the Deccan plateau. Early proposals such as the Kennedy plan in 1852 called for lines parallel to the coast in order to economise on costs. Some were never built because subsequent officials opted for more expensive routes cutting through mountains (Davidson, 1868). We return to the Kennedy plan below as an instrument for route placement.

Although British firms built the railways, the GOI dictated route placement. What guided their decisions? Military, commercial, and famine concerns were cited as the main drivers in official correspondence (Hurd, 1983). Following the Sikh Wars in the 1840s and the Indian Mutiny in 1857, the British were cognisant of the need to transport troops and supplies across the country at low cost. Existing transport routes were of poor quality and slow, which made it necessary to station troops at multiple locations in the event of an uprising (Parliamentary Papers, 1854). On the commercial side, British merchants lobbied for Indian railways that would connect the ports to cotton-growing regions in the interior, and from the eastern and western ports to Delhi in the north. Another consideration was famines. Following devastating crop failures and famines in the 1870s, the GOI built "protective lines" in famine-prone regions of the South. Finally, a few small lines were built connecting hill stations such as Simla where British officials liked to spend their summers. While not random, the network of railroads across districts was not uniformly indicative of positive or negative selection. Rather, a mix of factors affected where and when railroads were built. Coastal districts with important ports were connected early as were those in the Ganga valley. Yet, a few cotton-growing interior districts were connected before 1881, as were districts closer to Afghanistan. Neither group would be considered positively selected for rail access. To address the endogeneity of railroads, we compare cohorts within districts in panel models and use an instrumental variables strategy among other cross-sectional models.

3. Data

We construct a new district-level dataset for British India from 1881 to 1921 in order to test the relationship between railroads and education. Our data pulls information from four primary sources:

(1) the decennial censuses of 1881 to 1921, which we use to measure literacy and several other control variables;

(2) the 1934 edition of *History of Indian Railways Constructed and in Progress*, which we use to code the opening dates of the railroad;
(3) the District Gazetteers of India, which we use to code primary and secondary enrolment rates, and;
(4) multiple sources of Geographic Information System (GIS) data, which we use to construct geographic controls.

We begin with the 1881 census rather than the first 1872 census because of incomplete coverage and inconsistencies in the 1872 census (Walby and Haan 2012).[5] We also focus on British Indian districts because the Princely State data are inaccurate in the early censuses.[6] We describe our data sources in the remainder of this section.

3.1. **Measuring Literacy.** The colonial census reports detailed literacy by gender and age-cohort. From 1901, the census also reports English literacy. Despite its richness, enumerating literacy over time is difficult because of changes in definition and measurement. In the 1881 and 1891 censuses, individuals were classified into three categories: literate, learning, and illiterate. Yet, enumerators were given no guidance on measuring literacy or accounting for learners apart from an age threshold (Gait 1913).[7] Age bins in these early censuses were also different across provinces.

Beginning with the 1901 census, the "learning" category was dropped and literacy was reported for standard age bins: those under age 10, aged 10 to 15, aged 15 to 20 and those over age 20. A uniform definition was adopted, namely "the ability to read and write." But again, census administrators were not given official guidance on measuring literacy. This led to differences across provinces. Some used a rigorous standard while others enumerated individuals as literate if they could sign their name (Gait 1913).[8] It was only in 1911 that a uniform standard, the ability to read and write a short letter, was introduced. This makes literacy in the 1901 and later censuses difficult to compare. For example, many children under age 10 were counted as learners in the 1891 census, then some children under 10 were recorded as literate in the 1901 census, but not in subsequent censuses. Indeed, we observe literacy decline in some districts for those under age 10 between 1901 and 1911 because of these problems.

[5] Walby and Haan (2012) summarise the many issues with the first 1872 census including incomplete territorial coverage and inconsistent enumeration across provinces. Walby and Hann aptly quote official opinion: "Later commentators said that the only consistency in the 1871-72 Indian census was the "uniform absence of uniformity."" (p. 309)

[6] Colonial India encompassed British India with territories that were under direct British rule and Princely States that were governed by Indian rulers. Territorial coverage of the Princely States is incomplete and inconsistent up to 1911 (Census of India, 1901; 1911).

[7] See the education chapter in the Census of 1911, Part 1 - Report written by E.A. Gait.

[8] Gait writing in the 1911 census attests to the inconsistency: "In some parts criteria similar to those mentioned above appear to have been taken, while in others persons were entered as literate who could do little more than write their own name and spell out a few simple printed words." (p. 291)

To get around these issues, we focus on cohort literacy in the 1911 and 1921 censuses, when literacy was uniformly measured. The cohorts reported in these censuses are under 10, 10-15, 15-20, 20 and above. Using cohort literacy allows us to draw on the more accurate 1911 and 1921 census while retaining the ability to exploit the panel-like changes across cohorts over time within a given district. These censuses exploit more intensive margin variation in exposure to railroads. To exploit the more extensive margin growth of railroads before 1901, we perform two additional analyses. The first analysis uses the panel-like variation across cohorts within the 1901 census. The second relies on cross-sectional regressions estimated separately for each census year from 1881 to 1921. We address the endogeneity of railroad exposure in these cross-sectional exercises using instrumental variables, matching, and grid cell fixed effects. We describe these methods in the next section.

Table 1 shows literacy by cohort, gender, and language from 1901 to 1921. These are crude literacy rates equal to the number of literates in each group divided by the population of that group. Focusing on the cohort 20 and above, total literacy increased from 6.45% in 1901 to 8.41% in 1921. Men were more literate than women, though this gender gap narrowed over time. Men were 17 times more likely to be literate than women in 1901 compared to 9 times more likely to be literate in 1921. English literacy was low in absolute terms, but was a sizeable share of all literacy. Almost 15% of literate individuals in 1921 were, for example, also literate in English. Most children typically learned to read and write in a vernacular language before learning English (Sharp, 1918). So: English literacy was, in particular, a measure of upper-tail human capital (Basu, 1974).

Table 2 shows total, male, and female literacy for each cross-section from 1881 to 1921. As noted above, we do not have comparable data on cohort or English literacy earlier in the nineteenth century. Total literacy doubled from an average of 3% in 1881 to 6% in 1921. Differences by gender are again visible. Moreover, the standard deviation and range highlight the large differences across districts. Figures 2 through 4 show the distribution of total, male and female literacy across districts in each census year. While the distribution of literacy was highly skewed in 1881, it became more dispersed by 1921. Less than 1 person in 10 could read and write in over 95% of districts in 1881. By 1921 more than 1 person in 10 was literate in roughly 10% of districts, with a maximum literacy of 32% in Madras city. Male literacy was more dispersed than total literacy, as shown in Figure 3. It increased on average and became more dispersed from 1881 to 1921. Female literacy increased from its low base in 1881, yet the distribution remained highly skewed as late as 1921, as shown in Figure 4. Figure 5 shows literacy maps by quintile for 1881, 1901 and 1921.

3.2. **Measuring School Enrolment.** Unlike literacy, which measures the stock of human capital, enrolments capture the flow of human capital. As we expect railroads to affect the stock of literacy by increasing the flow of children in school, we complement our analysis of literacy with an analysis of school enrolment.

District enrolments are not reported in the colonial census, or national reports. Rather, they are reported in many district gazetteers. These sources are less uniform than the decennial census. Nonetheless, we construct a series on primary and secondary enrolment rates between 1894 and 1911, years with the most uniform data. This is an unbalanced panel, as most provinces report enrolment for a subset of years with only a few provinces reporting more years.

Primary school enrolment is recorded as the number of children enrolled in primary schools divided by the cohort under age 10. Secondary enrolment is children in schools other than primary schools, divided by the cohort aged 10 to 15.[9] Many secondary schools had attached primary classes, so some primary aged children will, then, be included in secondary enrolment. Such primary classes were of higher quality than regular vernacular primary schools (Sharp, 1918). Although this will introduce measurement error in both primary and secondary enrolment, we are unaware of district-level enrolment data of all primary school children, regardless of school type. Our measure of primary enrolment averaged 4.3%, compared to 2.4% for secondary enrolment. These enrolment rates are highly correlated with literacy, as we would expect. For instance, the correlation between 1901 literacy and total enrolment is 0.88, primary enrolment is 0.81 and secondary enrolment is 0.6.

3.3. **Measuring Access to Railroads.** To estimate the effect of railroads, we follow Fenske, Kala and Wei (2020) to code the opening dates of railway access in each district. Fenske, Kala and Wei (2020), following a procedure similar to Donaldson (2018), construct a polyline shapefile of the Indian railway network with an opening date for each segment. These dates are based on the 1934 edition of *History of Indian railroads Constructed and In Progress*. For each listed railway line, they record the opening dates along with the beginning and end points of each line. We intersect this shapefile of railway lines with a map of modern sub-districts. Using a GIS mapping of colonial districts to these modern sub-districts, we compute the earliest year that each colonial district is connected to the railroad.

Two common methods of measuring railroad access are (1) simple indicators for whether a location is connected to a railroad or not (e.g. Atack et al., 2010; Andrabi and Kuehlwein, 2010) and (2) market access (e.g. Donaldson and Hornbeck, 2016). Neither of these measures is well suited to our question. Both measures fail to capture whether exposure occurs over the typical ages of elementary schooling for members of the population. Nor do they capture the duration of this exposure. Market access is derived from structural trade models and

[9]By definition, secondary enrolment includes students in colleges and other schools. In Bengal where we have detailed enrolment information, high school and middle school enrolment accounts for 77% of secondary enrolment. In less advanced provinces, this percentage is likely to be higher because there were fewer colleges, training and other schools as there were in Bengal.

summarises well the economic forces acting on outcomes such as city growth and agricultural incomes. It is less suited to other potential mechanisms linking railroads to literacy, such as the diffusion of cultural practices and of information. We thus, in our synthetic panel analysis, use two measures of the cumulative number of years that a given cohort in a given district was exposed to railroads up to a cohort's elementary school ages. We also show robustness checks using an indicator for railroads and a market access variable in the cross-section.

The census reports literacy for four cohorts: under 10, 10-15, 15-20 and over 20. We define two alternative measures of railroad years assuming railroads affect literacy only up to the beginning or end of elementary school. This is a reasonable assumption because most children are literate or not by the end of elementary school (age 12).[10] If, for example, a railroad arrived in a district in 1893, it would not affect literacy for the cohort 20 and above in 1901 because they would be age 12 and above in 1893, and so would have finished primary school. If railroads arrive after the youngest age in a cohort is out of elementary school, we assume this cohort has no exposure to railroads (coded as 0). Unlike the 20 and above cohort, the arrival of railroads in 1893 would affect cohorts under age 10 and 10-15 in 1901 because they would presumably be in elementary school as railroads arrive. Our railroad measure captures such differences across cohorts.

Since the age bins do not perfectly correspond to elementary school years, we use the youngest age in the bin to measure cumulative exposure in elementary school and construct two measures of railroad exposure. Our first measure is the number of years a railroad has been operating in a district minus the number of years since the youngest member of a cohort would have regularly begun elementary school, i.e. at age 6.[11] Denote this number of years since schooling began as $y(c)$. For cohorts aged 20 and above, $y(c)$ is 14. For cohorts aged 15-20, it is 9. For cohorts aged 10-15, it is 4. For cohorts aged below 10, it is 0. For cohort c, $y(c)$ years since schooling began, in district d, with a railroad that opened in year r, measured in census year t, we define our treatment measure $RailroadYears_{cdt}$ as:

$$RailroadYears_{cdt} = \begin{cases} \max\{t - r - y(c), 0\} & \text{if } r \leq t, \\ 0 & \text{if } r > t. \end{cases} \tag{1}$$

Our second measure is the number of years a railroad has operated in a district minus the number of years since the youngest member of a cohort would have regularly finished elementary school. This measure assumes railroads affect literacy up to age 12 for the index

[10] See Sharp (1918) and other official Quinquennial Reviews of Education for discussion on colonial primary schools, enrolment and literacy.

[11] We use age 6 as the beginning of elementary school because the Indian compulsory school schemes in the 1910s used age 6 as an entry point. Sharp (1918) notes primary school lasted for six years and could begin as early as age 5.

age in a cohort. In equation (1), this is equivalent to replacing $y(c)$ with 8 for the cohort age 20 and above, 3 for cohorts aged 15 to 20, and zero for the cohorts aged 10-15 and 0-10.

As constructed, the two measures bound the duration of exposure of railroads. To make these measures more concrete, consider the following hypothetical. Suppose a railroad opens in a district in 1901. In the 1911 census, the railroad has been active for 10 years. In the cohort aged 10-15, the youngest member of that cohort was 6 in 1907 and has not yet turned 12. By our first measure, the railroad exposure for that cohort is $\max\{1911-1901-4, 0\} = 6$ years. By our second measure, the railroad exposure for that cohort is $\max\{1911 - 1901 - 0, 0\} = 10$ years.

To give a real example, Dehra Dun district was connected to a railroad in 1900. In the 1911 census, an individual aged 20 would have been 9 when railroads arrived in 1900. Our first measure codes the cohort 20 and above as untreated (0) in 1911, while our second measure codes this cohort's exposure as 3 years in 1911. Our first measure assumes parents decide whether to enrol their children in school based on cumulative exposure to the railroad up to the beginning of elementary school (age 6). Our second measure assumes parents decide whether to keep their children in elementary school based on cumulative exposure to the railroad up to the end of elementary school.

One may be concerned that age was incorrectly reported to census enumerators, which could introduce measurement error in cohort literacy. Indeed, age heaping at even numbers and multiples of five was common in colonial India (Census of India, 1911). But, census enumerators estimated an individual's age if it was at odds with their appearance. Census officials believed the age enumeration by cohort was reasonably accurate although the number of people at a specific age say 2 years old may be incorrect.[12] Using two measures of railroad exposure further alleviates concerns of measurement error as does the cross-sectional analysis on total literacy.

For the cross-sectional analysis, we count the number of years a district has been connected to a railroad in each census. Tables 1 and 2 report summary statistics on our various measures of railroad access. As seen in Table 2, 50% of districts were connected to a railroad by 1881, increasing to 96% by 1921. Indeed, most of the increase happened before 1901, with 87% of districts already connected. The railroad years measure better illustrates the variation across districts. For example, the number of railroad years averaged 7.4 years across districts in 1881, increasing to 22 years in 1901, and 41 years by 1921. Comparing the railway and

[12]According to PJ Mead and G Laird Macgregor in the 1911 Census of India for Bombay, "the census is taken on each occassion by the same class of individual dealing with much the same sort of material, and with the vast numbers that form our population the errors tend to counteract each other and age returns *en masse* are probably much nearer the mark that they appear to be, though the precise number at any particular age period is probably quite inaccurate" (Census of India, p. 75).

literacy maps suggests a small positive correlation between railroads and literacy. Figures 6 and 7 confirm the positive correlation for 1881 and 1921.

3.4. **Geographic and Socioeconomic Controls.** India has a wide range of terrain with the Himalayas in the north, mountain ranges along the east and west coasts, the Thar desert in Rajasthan, alluvial plains along the Indus and Ganga river valleys, and the Deccan plateau. Such differences in topography affected the railroad network because of the inherent difficulty in building railroads crossing mountains and deserts. These differences may also be relevant to explaining literacy gaps between districts. To this end, we construct a rich set of geographic variables in order to control for the selection into railroad exposure driven by geography.

In particular, we collect data on the latitude and longitude coordinates of the centroid of the district, which we compute ourselves. We control for ruggedness from Nunn and Puga (2012). We control for altitude, precipitation, temperature, slope, and suitability for growing specific crops such as cotton, dryland rice, wetland rice, and wheat, averaged over raster cells within a district. These are taken from the Food and Agriculture Organization of the United Nations' Global Agro-Ecological Zones (FAO-GAEZ) data portal. While the FAO data are based on current conditions, they are exogenous to human action and represent expected yields under low-input rainfed agriculture, and so have become widely used, including in economic history (e.g. Dimico et al., 2017; Kung and Ma, 2014). Since proximity to the coast and rivers likely influenced railroad access, we include indicators for rivers and coastal districts as captured in Natural Earth Data's shapefile maps of rivers and coasts. We also control for medieval ports recorded by Jha (2013). We control for the seasonality of rainfall. In particular, using data on historic rainfall from Willmott and Matsuura (2018), we compute the Feng et al. (2013) entropy-based measure of seasonality. Finally, we control for the Kiszewski et al. (2005) index of the stability of malaria transmission.

We also control for the religious and caste composition of a district. In particular, we include the share of Brahmans, traditional Hindu elites in the caste hierarchy that enjoyed higher literacy on account of their traditional occupations as priests and teachers. We also include the shares of Muslims, Christians, and tribal groups. Such shares are intended to capture historical differences in education among these groups that may be correlated with railroad access.[13] These data are taken from the colonial censuses.

[13]Some may view these as bad controls because these groups did not settle randomly across India. Our results are similar, if anything showing larger effects for railroads, when we do not control for them.

4. Estimation Strategy

Our main results exploit variation within districts and across cohorts to identify the effects of railroads on literacy. We complement this synthetic panel exercise with cross-sectional results using instrumental variables, grid cell fixed effects, and matching techniques.

4.1. **Synthetic Panel.** We estimate the following model using the log of the literacy rate by year, district, and cohort as the outcome.

$$\ln(LiteracyRate_{cdt}) = \beta RailroadYears_{cdt} + \theta_d + \delta_p \times \eta_t + \delta_p \times \gamma_c + \epsilon_{cdt} \tag{2}$$

In this model, $LiteracyRate_{cdt}$ is literacy for cohort c in district d and census year t. We transform literacy into logs because it is a highly skewed variable, as shown in Figure 5. We estimate the model for $t \in \{1911, 1921\}$ and cohort $c \in \{0-10, 10-15, 15-20, 20+\}$. $RailroadYears_{cdt}$ measures the cumulative years of railroad exposure for cohort c in district d in year t.

We control for several fixed effects. First, district fixed effects, θ_d, capture unobservable time-invariant district features that lead some districts to get railroads before others and that may correlate with literacy. Second, we include interactions of province $\times$ year and province $\times$ cohort fixed effects captured by $\delta_p \times \eta_t$ and $\delta_p \times \gamma_c$ to control for provincial changes in census enumeration methods by year and cohort. We do not include separate fixed effects for year (η_t) and cohort (γ_c) because they are collinear with $\delta_p \times \eta_t$ and $\delta_p \times \gamma_c$. Official guidance on census enumeration was set by provinces, administrative units larger than the district (Gait 1913). Such flexible controls address most measurement concerns related to literacy as well as accounting flexibly for omitted variables at the province and cohort level that may change over time.[14] We cluster standard errors by district to account for serial correlation over time. And, as a robustness check, we estimate Conley (1999) standard errors that account for spatial correlation with cutoffs ranging from 200km to 500km.

In this setup, we identify the effects of railroads using variation in cumulative exposure to railroad years across cohorts within districts over time. The key identifying assumption is that such exposure in railroad years is uncorrelated with the error term ϵ_{cdt}. We believe this is a reasonable assumption given the flexible fixed effects included in the model. As a robustness check, we run the same analysis using the 1901 census and controlling for district fixed effects and province $\times$ cohort fixed effects.[15] Since we use only the 1901 census for this exercise, changes in the standards used to measure literacy in different censuses are not an issue. We report other robustness checks in the results section.

[14] With four cohorts and two years, we do not have sufficient power to include district-cohort and district-year interactions, akin to a triple difference specification.

[15] We are unable to do the same exercise for 1881 and 1891 because they do not report uniform age bins.

4.2. **Cross-Section.** We complement the panel methods with several cross-sectional models as follows.

4.2.1. *Ordinary Least Squares.* We exploit the complete data from 1881 to 1921 using repeated cross-sections. For each census year, we estimate a separate OLS regression of the following form:

$$\ln(LiteracyRate_{dt}) = \beta RailroadYears_{dt} + \gamma' x_{dt} + \delta_p + \epsilon_{dt} \tag{3}$$

We estimate this regression separately for $t \in \{1881, 1891, 1901, 1911, 1921\}$. In this equation, $\ln(LiteracyRate_{dt})$ is the log literacy rate in district d in year t. $RailroadYears_{dt}$ is the number of years district d in year t has had a railroad. This is 0 if the district is unconnected in t. We do not adjust $RailroadYears$ because the outcomes are total, male, female, and English literacy. The vector x_{dt} includes the GIS and social controls described in section 3. We also include province fixed effects captured by δ_p. Finally, ϵ_{dt} is the error term. We estimate robust standard errors and report results with Conley standard errors with distance cutoffs ranging from 200km to 500km as a robustness check.

Such a regression may generate biased estimates of the causal effect of railroads. For example, if more developed districts were the first to receive railroads, our estimate of railway years would be biased up because it would conflate the effect of railroads with those of prior development. On the other hand, if famine-prone areas received access early on, then our estimates would likely have a negative bias. Indeed, military strategy also does not provide clear guidance on the potential selection problem. Railroads from Calcutta to Delhi facilitated quick movement of troops, but were also of immense commercial value in transporting goods. This would suggest positive selection. Railroads from Delhi heading northwest towards Afghanistan were of less commercial value and would, alternatively, suggest negative selection. To address such endogeneity concerns, we employ matching, fixed effects, and instrumental variables solutions.

4.2.2. *Grid Cell Fixed Effects, Matching and Market Access.* We complement the OLS results with grid cell fixed effects and matching models. For the grid cell fixed effects, we construct $2° \times 2°$ grid cells based on the the latitude and longitude coordinates of district centroids. 1° of latitude is roughly 111 kilometers. 1° of longitude ranges from roughly 111 kilometers at the equator to 85 kilometers at $40°N$ latitude, just south of New York, Madrid, or Tashkent. These grid cells will range, then, between 10,000 and 12,000 square kilometers in area in our data. For comparison, modern-day Tripura state has an area of 10,486 square kilometers, while Punjab has an area of 50,362 square kilometers. We include fixed effects for these grid cells in the OLS regressions. This ensures we are comparing neighbouring districts with different durations of railroad exposure for identification. We include our standard controls in these specifications.

We also estimate nearest neighbour matching models using the rich GIS controls to match districts on their time-invariant characteristics. Since 87% of districts are connected by 1901, we focus on the 1881 and 1891 cross-sections, in which there is more balance between connected and unconnected districts. In these nearest neighbour matching models we measure railroad connection as an indicator variable. Apart from railroad years and an indicator for railroad access, we calculate a district-level market access variable following Donaldson and Hornbeck (2016). Eaton and Kotrum (2002) discuss different θ values for this calculation. We report results using θ of 8.22. That said, our results are robust to other θ values.[16]

4.2.3. *Instrumental Variables: 1852 Kennedy Plan.* Finally, we construct an instrument for $RailroadYears_{dt}$ using Major J. P. Kennedy's 1852 proposal for building railroads in India. Major Kennedy was the Consulting Engineer for the GOI and played a key role in planning India's railway network. He pushed for building low-cost railroads that, in his view, would confer innumerable benefits. As stated in his own words:

> It is not sufficient to be convinced as I am, that the establishment of Railways in India is an essential preliminary to the attainment of the highest degree of efficiency of which our military and civil administrations are capable; to the prevention of local famine, and to the uniform dispersion of food; to any vigour and activity in manufacture or commerce; to the increased consumption of English goods: to the power of competing with America in furnishing to England raw cotton and other important articles: in short, to the growth of everything connected with the extension of British interests in India as well as with the industry, the wealth, and the comfort of its vast population (Parliamentary Papers 1854, p.3).

Yet, Major Kennedy was aware of the costs of building railroads. So he emphasised lower-cost routes connecting the ports with the interior. In particular, his plan called for a network in "strict harmony with the natural advantages" of the country. Unlike routes that would cut through the Eastern and Western coastal ranges of India, his plan called for routes that favoured softer gradients, following the coast and natural topography.

Donaldson (2018) used portions of the Kennedy plan that were not implemented to construct placebo lines. In many cases, however, Kennedy's routes were adopted, as seen by comparing the Kennedy plan in Figure 8 to the actual network in Figure 1. In other cases, however, more expensive routes were selected. Exploiting geographical features in favour of low cost routes was Kennedy's focus. Hence, we are assuming here that, conditional on controls, the 1852 Kennedy plan is uncorrelated with factors that would affect literacy other than through access to railroads. To construct the instrument, we convert the map

[16]We have estimated market access for θ values of 1, 3.73, 8.22, 12.86 and 26.83.

of Kennedy's proposal into a polyline shapefile. We then calculate the shortest distance of each district from this route. We use the log of (one plus) distance to the Kennedy plan as an instrument for $RailroadYears_{dt}$.

We also report results using two other instruments using a strategy common in the transportation literature – constructing a minimum spanning tree (Morten and Oliveira, 2016; Faber, 2014; Jedwab and Moradi, 2016; Fenske, Kala and Wei, 2020). First, we build a tree spanning Indian cities with populations greater than 10,000 in 1850 using the set of cities recorded in Chandler and Fox (1974). We choose 1850 because it precedes the construction of railroads. Using Prim's algorithm, we construct the shortest tree that spans these 97 cities. Figure 9 shows a map of this tree superimposed on the 1881 railway network. After constructing the tree, we compute the distance of each district from the spanning tree. We then use the log of (one plus) distance to this tree as an alternative instrument for $RailroadYears_{dt}$.

Second, we build a tree spanning the 54 British military cantonments that existed as of 1864. If military concerns drove the placement of railroads, we expect military cantonment towns to get early access. Following the same approach using Prim's algorithm, we compute the distance from the district to the tree spanning the military cantonments. We then use the log of (one plus) distance to this military cantonment tree as an instrument for $RailroadYears_{dt}$. Figure 9 shows a map of this spanning tree.

In our view, these instruments are inferior to the Kennedy plan instrument. Apart from access to railroads, cities may attract educated people and offer higher returns to education. Such channels could violate the exclusion restriction. The same logic applies to military cantonments where British officers lived. Such critiques are less applicable to the Kennedy plan instrument because it relies more on geographic differences favouring easier terrain for building railroads. As robustness checks, we report results using all three instruments in the same regression along with an over-identification test and results of each instrument used individually.

5. Results

5.1. Synthetic Panel. Table 3 shows our main results on railroad exposure, which exploit variation across cohorts within districts in 1911 and 1921. Column (1) focuses on the log of total literacy, column (2) on male literacy, and column (3) on female literacy. In the second panel we show results for English literacy. We report results for non-English literacy in the bottom panel. We calculate non-English literacy by subtracting English literates from total literates and dividing by the relevant population.[17]

[17]Appendix Table A1 shows regressions using spatially adjusted Conley (1999) standard errors. In particular, we use the implementation for panel data developed by Hsiang (2010) and Fetzer (2019), and report results

As seen in Table 3, the coefficient on railroad exposure is positive and significant for total and male literacy, but not female literacy. In terms of magnitude, the standardised β coefficients (multiplying the β coefficient in Table 3 by the standard deviation of cohort railroad years, 17 years, and dividing by the relevant standard deviation of log literacy) range from 0.29 standard deviations for total and 0.31 standard deviations for male literacy in the top panel. We find smaller effects on male English compared to male non-English literacy with standardised coefficients at 0.25 for English and 0.34 for non-English literacy. Unlike for males, we find small and insignificant effects of railroads on female literacy, female English and female non-English literacy. These gender differences make sense in context; few Indian women could read and write, with cultural norms often prohibiting girls from attending school outside the home (Nurullah and Naik, 1951). It is thus unsurprising then that railroads did not increase female literacy.[18]

This exercise includes the cohort age 20 and above. Individuals aged 20 in this cohort began elementary school fourteen years ago at age 6, but others at age 30 in the cohort were past elementary school fourteen years ago when they were age 16. To ensure our results are not driven by such mismeasurement in cohort railroad years, we estimate the same regressions as above for the cohorts under 10, 10-15 and 15-20, removing those aged 20 and above from the sample. Any measurement error in cohort railroad exposure is smaller for these tighter age bins. As seen in Table A4, the results are similar, albeit with stronger results for non-English literacy compared to English literacy.

As outlined in section 3, our first measure of cohort railroad exposure uses an index age in a cohort based on the beginning of elementary school at age 6. Our second method of constructing cohort exposure uses an index age for a cohort based on the completion of elementary school at age 12. Table A5 shows the results using these different exposure measures. We find similar results with positive and significant effects only for male literacy. In terms of magnitude, they are marginally smaller at 0.26 standard deviations for total literacy and 0.29 standard deviations for male literacy.

An advantage to using the 1911 and 1921 census is the consistent enumeration of literacy across the two years. A disadvantage is that 94% of districts are connected by 1911. Unlike Table 3, we exploit across cohort within district variation using the 1901 census in Table A6. We find positive effects of railroads for male and English literacy, although the estimates on total male and non-English male literacy are smaller in magnitude and less precisely estimated than for 1911 and 1921. Increasing railroad exposure by 14 years (the standard

with a maximum of five lags and distance cutoffs of 200 kilometers. We find the same results for distance cutoffs of 300 and 500 kilometres.

[18] Our results are robust to dropping the four cities of Bombay, Calcutta, Delhi and Madras as seen in Table A2. They were among the first cities to be connected to a railroad. The results are also not driven by a specific province. Table A3 shows they are robust to dropping one province at a time.

deviation on cohort railroad years) increases male literacy by 0.09 SD and male English literacy by 0.24 SD. We again find small and insignificant effects on female literacy.

Taken together, these results suggest railroads had large and positive effects on male English and non-English literacy. These results are not driven by mis-measurement of exposure in the cohort age 20 and above. They are similar across different measures of cohort exposure, across more intensely and less intensely treated cohorts in 1911-1921, and across cohorts within districts in 1901.

5.2. **Cross Section.**

5.2.1. *Ordinary Least Squares.* We turn next to cross-sectional results. Table 4 reports OLS estimates for each census year. While we report robust standard errors in this table, we show in Table A7 that our results are similar when we use Conley (1999) standard errors to adjust for spatial correlation in the error term. Columns (1) to (3) show results for log literacy with no controls in (1), including province fixed effects in (2), and including province fixed effects with the full set of controls in (3). In columns (4), (5) and (6) we report results for male, female and English literacy. Two patterns stand out. First, the estimates are positive and significant across specifications. Second, the effects are larger for female and English literacy compared to male literacy.

Standardised β coefficients range from 0.15 to 0.32 standard deviations, with those for English and female literacy being on the higher end of the range. For example, in column (3) for 1881, a one standard deviation increase in railroad years translates into a 0.21 standard deviation increase in literacy. By 1921 the standardised magnitude decreases slightly to 0.18. Unlike these small changes in effect sizes over time, we find bigger differences by gender. For example, the 1921 standardised coefficient for female literacy at 0.27 is significantly larger than for male literacy at 0.15. Nonetheless, we are cautious drawing strong conclusions from these results on female literacy, because we find small and insignificant effects of railroads on female literacy in panel regressions that exploit variation across cohorts within districts in the same census year.

5.2.2. *Grid Cell Fixed Effects, Matching and Market Access.* Table A8 presents results comparing neighbouring districts using grid cell fixed effects. We first construct a $2° \times 2°$ grid using the latitude and longitude coordinates of each district's centroid. Then we include grid cell fixed effects along with the geographic and social controls. The coefficients on railroad years are positive and significant with similar effect sizes to the OLS estimates. Comparing neighbouring districts does not attenuate the coefficients, which suggests our rich set of controls capture many of the relevant differences across districts.[19]

[19]We also ran robustness checks with grid cell fixed effects in our synthetic panel. We find similar effects with positive and significant effects of railroads on total, male, female and English literacy.

Because we have these rich controls, we also undertake a matching exercise for the 1881 and 1891 cross-sections. In these years we observe both connected and unconnected districts. Table A9 reports average treatment on the treated (ATT) estimates using a nearest neighbour matching exercise. We measure railroad exposure as an indicator variable and match districts with railroads to those without using the geographic and social controls. Again, railroads positively and significantly predict literacy. Access to railroads increases 1881 male literacy by 20% and female literacy by 50%.

Finally, we report results using a market access variable in Table A10, which show positive and significant effects of market access on literacy. Unlike railroad years, market access is not significantly correlated with female literacy in the 19th century but becomes significant in the 20th century. Since this variable does not have units, we cannot interpret it like the railroad indicator or railroad years measure.

5.2.3. *Instrumental Variables.* Table 5 shows second stage instrumental variables results using the instrument based on the Kennedy plan. We show the first stage results linking log distance from the Kennedy Plan to railroad years for each year in Table A11. As described in Section 4, this plan was developed in 1852, before the construction of railroads had begun. It proposed low-cost railroads favouring areas with mild gradients. Columns (1)-(6) correspond to the same outcomes and controls as in Table 4. As seen by the large Kleibergen-Paap F-statistic (KPF), the Kennedy plan strongly predicts railroad years in each census year. Our IV results confirm our previous findings: railroads positively predict literacy.

In the earlier years, the IV estimates are larger in magnitude for females than males. In standardised terms, the effects of railroad years on 1881 female literacy are 0.27 standard deviations compared to the OLS estimate of 0.32 in Table 4. By 1921, the estimates are significant for male literacy but with stronger effects for female and English literacy. Unlike 1881, the magnitude of the effects is larger in 1921, ranging from 0.37 standard deviations for male literacy to 0.79 standard deviations for female literacy. These IV estimates are local average treatment effects (LATE), namely the effect of increasing railroad years for those districts that gained access to railroads earlier because of their proximity to the 1852 Kennedy plan. This translates into more isolated districts incidentally being connected to a railroad because they are on a direct line between major centres. Most such connections came after the early direct connections from Delhi in the interior to the ports (Hurd, 1983). Such isolated places may have benefited more from railroads, which would account for the larger results in 1921.

Table 6 shows the results using the three instruments together: (1) log distance from 1852 Kennedy Plan; (2) log distance from a tree spanning military cantonments circa 1864, and; (3) log distance from a tree spanning cities circa 1850, before railroad construction began. We find large and positive effects of railroads using these three instruments together. The

KPF statistic is larger than for the single Kennedy Plan instrument. Moreover, we are unable to reject the Hansen over-identification test across the specifications. In Table A12 we show the IV results separately for the (1) distance from a tree spanning military cantonments circa 1864, and (2) distance from a tree spanning cities circa 1850, before railroad construction began. The IV results point to uniformly large, positive and significant effects of railroads. The effect sizes are similar over time, with larger magnitudes for female compared to male literacy.

Both the cross-sectional and synthetic panel methods, then, point in the same direction of positive and significant effects of railroads on male and English literacy. Unlike in the panel models, we find positive and significant effects on female literacy across the cross-sectional models.

5.3. **Discussion.** Are these effects big or small? To answer this question, we first benchmark our results against those in Atack, Margo and Perlman (2012). They estimate the effect of railroads on individual school enrolment in the United States. Their estimates suggest that increasing rail access across US counties in the 1850s predicts 56% of the increase in mean school enrolment between 1850 and 1860 (p. 16). We find smaller effects for India. In our case, increasing exposure to railroads between 1881 and 1891 predicts 28% of the actual increase in literacy.[20] It may well be infrastructure expansions have larger spillovers in more developed countries where schools were more widespread than in India.

Another way to consider the size of these estimates is in comparison to supply interventions. Chaudhary (2010) finds it would have cost the colonial government roughly 3 rupees to make an additional person literate using causal estimates of public education spending on literacy. To construct a similar estimate for railroads, we have to monetise the increase in railway years. One crude approach is to use the change in capital outlay and working expenses between the relevant years, which we obtain from Bogart and Chaudhary (2016). This suggests an increase in railroad years between 1881 and 1891 of 6.28 years translates into 844,889,000 rupees. This increase predicts 28% of the increase in literacy between 1881 and 1891, translating into 540,836 additional literates. Converting this into per capita terms suggests a cost of 1,562 rupees to make one additional person literate. This is a simple, illustrative back of the envelope exercise. Railroads conferred many benefits on Indian society that are not captured here. What this exercise merely shows is that railroad effects on schooling would have had to be implausibly large to be a cost-effective strategy to increase mass education.

[20]In this calculation, we multiply the increase in railroad years of 6.28 between 1881 and 1891 with the 1881 OLS estimate on railroad years in Table 4, column (3), to predict the increase in literacy by 1891. We then compare this predicted increase to the actual increase in literacy.

6. Mechanisms

In this section, we begin by documenting the proximate mechanism through which railroads increased literacy – greater school enrolment. We then provide suggestive evidence on the deeper mechanisms linking railroads to schooling. Previous work has documented that transportation investments generate heterogeneous effects on schooling (Adukia et al., 2020). Connecting previously unconnected places can increase the opportunity cost of schooling by increasing agricultural incomes or wages of low-skilled workers. This suggests a possible negative effect of transportation infrastructure on schooling, which would work against our results. On the other hand, access to transportation can increase the returns to education by increasing the wages of more educated workers compared to less educated workers. Transportation can also increase household earnings and ease liquidity constraints, which would lead to more education if schooling is a normal good. We use formal mediation models to consider the mediating effects of different channels in the cross-sectional OLS and IV frameworks. Because we cannot construct cohort-specific measures of these possible mediators, we do not use the synthetic panel framework for this exercise.

6.1. **Enrolment.** Table 7 shows the results on enrolment for the panel and cross-sectional methods. As seen in the top panel, where we include district and year fixed effects, increasing exposure to railroads has a positive and significant effect only on secondary enrolment. Indeed, the coefficient on primary enrolment is negative albeit insignificant. In terms of magnitude, a one standard deviation increase in railroad years increases secondary enrolment by 0.5 standard deviations in specification (5). We also find large and positive effects of railroads in the cross-sectional OLS and Kennedy IV models. The IV estimates are larger than the OLS estimates, with the effects decreasing between 1901 and 1911.[21] In comparison to literacy, these standardised β coefficients are larger for both the panel and cross-sectional models. This is unsurprising. We would expect bigger effects of railroads on the flow of children into school compared to the stock of literates.

Both the enrolment and literacy findings also point to larger effects on what, in the context of colonial India, constituted upper tail human capital. We find positive effects on secondary enrolment, which in our data sources includes higher quality primary classes and English instruction. With literacy, we find positive effects on English literacy, which was more common among Indian elites than the rest of the population (Basu, 1974). The benefits of railroads were thus concentrated, rather than shared by the general population.

6.2. **Agricultural Income and Land Taxes.** Before railroads, transportation in India was of poor quality, expensive, and unreliable (Hurd 1983). Railroads had a large effect on price convergence, trade, and agricultural income. According to Donaldson (2018), about

[21] We find similar results using grid cell fixed effects.

half the increase in agricultural income due to the railroad came from falling trade costs. Are rising agricultural incomes a mediator from railroads to higher literacy? Theoretically, an increase in agricultural income leads to both income and substitution effects. If schooling is a normal good, we would expect parents to "buy" more education, generating a positive effect on literacy. On the other hand, an increase in agricultural income may increase the low-skilled wage and the returns to child labor, generating a negative effect on literacy.

We study agricultural income as a possible mediator using Donaldson's (2018) series on rural income.[22] We begin with our instrumental variables framework following Dippel et al. (2020), who offer a formal mediation analysis nested in instrumental variables. This approach identifies a total effect from a treatment (in our case railroads) to an outcome (here, education) and then decomposes it into an indirect effect via the mediating factor (agricultural income in our case) and a direct effect from treatment to the outcome not via the mediator. According to Dippel et al. (2020), this procedure requires strong instruments in two first stage regressions to identify all three effects. The standard first stage regression from instrument to treatment generates a treatment first stage F-statistic, while another first stage regression from instrument and treatment to mediator generates another mediator first stage F-statistic.

Our treatment F-statistics are greater than 10, but we have very low mediator first stage F-statistics. That is, the instrument and treatment together are not strong predictors of the mediator. Although we present results from this exercise in Table A13, we are cautious in drawing strong conclusions because of our weak mediator first stage. The coefficients on total effects for enrolment and literacy are similar to our main IV results shown in Table 5. They are larger in magnitude for enrolment and smaller in magnitude for literacy.

In Tables 8 and 9, we conduct an alternative mediation analysis suitable for an OLS framework (Imai et al., 2010a; 2010b).[23] As is standard in these analyses, this method relies on the sequential ignorability assumptions that railroad years are quasi-randomly assigned, conditional on the geographic and social controls, and that the mediator is ignorable conditional on railroad years and controls. Tables 8 shows the mediation results for total literacy and Table 9 for secondary enrolment.[24] As seen in specifications (1) and (2) in the top panel, the coefficient on income is small, negative and insignificant.

In specifications (3) and (4), we also rule rule out a link from agricultural income to education via public funding. Surcharges on existing land taxes were a key funding source for rural district boards. Such boards managed rural primary education. While there could be a positive link in theory from railroads to agricultural income to land taxes, we find land

[22]Our results are the same if we use nominal income per area, real income, or real income per area.
[23]In particular, we use the implementation developed by Hicks and Tingley (2011).
[24]We find similar results for male and English literacy as seen in Tables A14 and A15. We focus on secondary enrolment because railroads did not effect primary enrolment as seen in Table 7.

taxes per capita are not a significant mediator for literacy or secondary enrolment. These results are unsurprising. Land taxes were fixed in nominal terms in eastern India in 1793 (Kumar, 1983). In these areas, land taxes were thus unconnected to late nineteenth and early twentieth century agricultural incomes. Even in other parts of the country, land taxes were revised infrequently, around every 30 years (Kumar, 1983).[25] There is no evidence that rising agricultural incomes mediate our railroad results.

We also collected data on unskilled wages to assess their relationship with railroads. Across fifty districts with reported data on agricultural wages, we find no significant correlation between the low-skilled agricultural wage and railroads in 1881 or 1891. If railroads did not increase the low-skilled wage, it is likely that substitution effects driven by rising incomes are not a significant channel from railroads to education.

Finally, we consider whether our results are mechanically driven by a larger supply of railway schools in districts with early exposure to railroads. Railway companies established separate schools for children of their European and Indian employees. The East Indian Railway Company was among the first to set up schools in places where there was sufficient demand among its employees.[26] Such discussions were just beginning in 1881, so they cannot account for the 1881 results between railroads and literacy. By 1911 there were 200 such schools in British India enrolling 7,500 children of which 42% were European.[27] To put in perspective, they account for 0.16% of total schools and 0.15% of total enrolment in 1911. They are too small to affect our results. We also directly test whether the presence of Europeans is mediating our results in the next section.

6.3. **Non-Agricultural Income and Returns to Education.** Apart from agricultural income, railroads may have increased non-agricultural income and the returns to education thus linking railroads to education. Unfortunately, there are no data sources that we know of that report wages by level of education or literacy. We indirectly test whether increasing returns to skill and rising non-agricultural incomes play a mediating role by looking at income tax revenues and the share of workers in industry, services and public administration (a subset of services).

Income taxes were assessed on non-agricultural income using a schedule that varied by income source. Salaries and pensions for example came under one schedule, while income from trade, commerce and professional employment came under another schedule (Alvaredo et al. 2017). Since income from agriculture was not taxed, this measure captures income from industrial and professional employment. The share of non-agricultural workers and

[25]We show the direct correlation between railroads and potential mediators in Table A16. While railroad years are correlated with agricultural income, they are uncorrelated with land taxes.

[26]Typically, parents were charged fees with some allowances for low income employees.

[27]Data from *Administration Report on the Railways in India for the Calendar Year 1911* (Govt of India, 1911).

income taxes are both proxies, then, for returns to education. Both measures conflate the supply and demand for educated labor. However, we believe they are decent proxies for returns. Of the two, income taxes are the better proxy, because the demand for educated labor was inelastic in colonial India. Increases in labor demand would thus lead to larger changes in wages rather than in the number of workers.

We use labor force data from Fenske, Gupta and Yuan (2020) that construct these measures using the decennial census. Income taxes on the other hand were reported in the district gazetteers for certain years with 1901 and 1911 being the most common reporting years. Specifications (5) and (6) in the top panel of Tables 8 and 9 show that income taxes have a positive and significant coefficient for both literacy and secondary enrolment. As seen in the top panel, income taxes mediate 40% to 50% of the effects of railroads on literacy, and 30% to 50% of the effects on secondary enrolment. Rising non-agricultural incomes may have led to income effects and eased liquidity constraints, leading more families to "buy" schooling for their children.

In the bottom panel of Tables 8 and 9, we look at the share of workers in industry, services and public administration. Service sector employment is a more significant mediator compared to industrial employment. It mediates anywhere from 14% to 25% of the effect of railroads on literacy and secondary enrolment. Lawyers and public administrators among other professionals were part of the service sector. Indeed, we confirm the mediating role of public administrators in the bottom panel of Tables 8 and 9. As seen, they account for a big part of the services effect. Such workers by definition were more educated than the rest of the population. And, these occupations also paid higher wages than other skilled occupations. These measures do, however, conflate income effects with rising returns to education. We have no way of disentangling these channels and interpret these results as evidence of their joint importance.

Tables A17 and A18 summarise results for other mediators we considered. The top panel shows the correlation between mediators and railroad years, while the bottom panel shows the OLS mediation analysis. Railroads carried both goods and people. So in theory they could have increased migration, which in turn may have increased literacy. However, we find no correlation between railroads and the share of migrants, or any evidence that migration was mediating the effect from railroads to literacy.[28] Indian railroads were built by the British GOI, so it may well be that districts with a larger share of Europeans had more exposure to railroads, which in turn could account for the positive effects on literacy. Yet again, we find no significant relationship between railroads and the share of Europeans, or on the mediating role of Europeans. Christians set up missions all over India and they may have chosen to settle in districts with easy access to railroads. Indeed, we find that railroad years

[28]Migrants are defined as people that are not born in their district of census enumeration.

are positively correlated with an indicator for whether a district had a Protestant mission as of 1911. But, these missions did not play a significant mediating role between railroads and literacy or secondary enrolment. In specification (4), we check whether railroad workers are driving the service sector result. Here we subtract railroad workers from the service sector. We find similar results for non-railroad service sector workers as total service sector workers.

7. Conclusion

We study the effects of railroads on Indian literacy and enrolment using district-level data from 1881 to 1921. We find positive and significant effects of railroads on male and English literacy. Our results are robust in both panel models where we exploit variation in railroad exposure across cohorts within districts and in cross-sectional models where we control for the endogeneity of railroad exposure using instrumental variables. Railroads lead to greater literacy via higher secondary enrolment. We find no evidence that agriculture is an important mediator. Rather, non-agricultural income and service sector employment are key mediators of the link between railroads and higher schooling. Railroads generated positive spillovers on education, but their effects were concentrated and not broadly shared.

References

(1) Adukia, Anjali, Sam Asher, Paul Novosad. (2020). "Educational Investment Response to Economic Opportunity: Evidence from Indian Road Construction." *American Economic Journal: Applied Economics* 12(1): 348-376.
(2) Atack Jeremy, F. Bateman, M. R. Haines, and Robert A. Margo. (2010). "Did Railroads Induce or Follow Economic Growth?: Urbanization and Population Growth in the American Midwest, 1850-1860." *Social Science History* 34(2):171-197.
(3) Atack Jeremy, M. Haines, and Robert A. Margo. (2011). "Railroads and the Rise of the Factory: Evidence for the United States, 1850-1870." In Rhode, P., Rosenbloom, J., and Weiman, D., editors, *Economic Evolution and Revolutions in Historical Time*, pp.162-179. Stanford University Press, Palo Alto, CA.
(4) Atack, Jeremy, Matt Jaremski and Peter Rousseau. (2014). "American Banking and the Transportation Revolution Before the Civil War." *Journal of Economic History* 74(4): 943-86.
(5) Atack Jeremy, Robert A. Margo and Elizabeth R. Perlman. (2012). "The Impact of Railroads on School Enrollment in Nineteenth Century America." Working paper: http://elisabethperlman.net/papers/PerlmanMargoAtack_SchoolRR_Draft8.pdf
(6) Banerjee, Abhijit and Esther Duflo. (2011). *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York: Public Affairs.

(7) Banerjee A, Duflo E, Qian N. (2020) "On the road: Access to Transportation Infrastructure and Economic Growth in China." *Journal of Development Economics* 145, (June): 102442.
(8) Baum-Snow N, Brandt L, Henderson JV, Turner MA, Zhang Q. (2017). "Roads, Railroads, and Decentralization of Chinese Cities." *Review of Economics and Statistics* 99(3):435-48.
(9) Bogart, Dan and Latika Chaudhary. (2016). "Railways in Colonial India: An Economic Achievement?" In *A New Economic History of India*, edited by Latika Chaudhary, Bishnupriya Gupta, Tirthankar Roy and Anand V. Swamy, 140-160. New York: Routledge.
(10) Chaudhary, Latika. (2016). "Caste, Colonialism and Schooling: Education in British India." *In A New Economic History of India*, edited by Latika Chaudhary, Bishnupriya Gupta, Tirthankar Roy and Anand V. Swamy, 161-178. New York: Routledge.
(11) Combes PP, Gobillon L, Zylberberg Y. (2020). "Urban Economics in a Historical Perspective: Recovering Data with Machine Learning." CEPR Discussion Paper No. DP15308.
(12) Conley, T. G. (1999). "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics*, 92(1), 1-45.
(13) Dimico, A., Isopi, A., & Olsson, O. (2017). "Origins of the Sicilian Mafia: The Market for Lemons." *The Journal of Economic History*, 77(4), 1083-1115.
(14) Donaldson, Dave and Richard Hornbeck. (2016). "Railroads and American Economic Growth: A "Market Access" Approach." *The Quarterly Journal of Economics* 131(2): 799-85.
(15) Donaldson, Dave. (2018). "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." *American Economic Review* 108 (4-5): 899-934.
(16) Faber, B. (2014). "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System." *Review of Economic Studies*, 81(3):1046–1070.
(17) Feng, X., Porporato, A., & Rodriguez-Iturbe, I. (2013). "Changes in Rainfall Seasonality in the Tropics." *Nature Climate Change*, 3(9): 811-815.
(18) Fenske, J., Gupta, B., & Yuan, S. (2020). "Demographic Shocks and Women's Labor Market Participation: Evidence from the 1918 Influenza Pandemic in India." CAGE Working Paper No. 494.
(19) Fenske, J, Kala N, Wei J. (2020). "Railways and Cities in India." CAGE Working Paper. 2020.

(20) Fetzer, T. (2020). “Can Workfare Programs Moderate Conflict? Evidence from India.” *Journal of the European Economic Association*, 18(6): 3337?3375.

(21) Fogel, R. (1964). *Railroads and American Economic Growth: Essays in Econometric History.* Baltimore: John Hopkins Press.

(22) Gait, E. A. (1913). *Census of India, 1911.* Calcutta, India: Office of the Superintendent of Government Printing.

(23) Great Britain (1929). *Indian Statutory Commission. Interim report of the Indian Statutory Commission. (Review of growth of education in British India by the Auxiliary committee appointed by the commission).* London: His Majesty’s Stationery Office.

(24) Hanlon, W. and Heblich, S. (2020). “History and Urban Economics.” *Regional Science and Urban Economics*, Forthcoming..

(25) Herranz-Loncán, A. (2014). “Transport Technology and Economic Expansion: the Growth Contribution of Railways in Latin America before 1914.” *Journal of Iberian and Latin American Economic History* 32(1): 13-45.

(26) Hicks R, Tingley D. Causal mediation analysis. The Stata Journal. 2011 Dec;11(4):605-19.

(27) Hsiang SM. Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America. Proceedings of the National Academy of sciences. 2010 Aug 31;107(35):15367-72.

(28) Hurd II, John. (1975). “Railways and the Expansion of Markets in India,” *Explorations in Economic History* 12(3): 263-88.

(29) Imai, Kosuke, Luke Keele and Dustin Tingley. (2010). “A General Approach to Causal Mediation Analysis.” *Psychological Methods*, 15(4) pp. 309-334.

(30) Jedwab, R. and Moradi, A. (2016). “The Permanent Effects of Transportation Revolutions in Poor Countries: Evidence from Africa.” *Review of Economics and Statistics*, 98(2):268–284.

(31) Kerr, Ian J. ed. (2001). *Railways in Modern India.* Oxford University Press, New Delhi.

(32) Kerr, Ian J. ed. (2007). *27 Down: New Departures in Indian Railway Studies.* Orient Longman, Hyderabad.

(33) Kumar, Dharma. (1983). “The Fiscal System” In The Cambridge Economic History of India, Volume 2: c. 1757-c. 1970, edited by Dharma Kumar, 905-944. Cambridge University Press: Cambridge.

(34) Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., & Sachs, J. (2004). “A Global Index Representing the Stability of Malaria Transmission.” *The American Journal of Tropical Medicine and Hygiene* 70(5): 486-498.

(35) Kung, J. K. S., & Ma, C. (2014). "Autarky and the Rise and Fall of Piracy in Ming China." *The Journal of Economic History* 74(2):509-534.

(36) McAlpin, Michelle Burge. (1975). "The Effects of Markets on Rural Income Distribution in Nineteenth Century India." *Explorations in Economic History* 12 (3): 289-302.

(37) Morten, M. and Oliveira, J. (2016). "The Effects of Roads on Trade and Migration: Evidence from a Planned Capital City." NBER Working Paper No. 22158.

(38) Nurulla, S. and J P. Naik (1951). *A History of Education in India: (during the British Period).* Bombay: Macmillan.

(39) Parliamentary Papers. (1854) 131. Railways (India). *Copy of memorandum by Major Kennedy, on the question of a general system of railways for India, referred to in the minute by the Governor-General in Council of 20 April 1853, relative to railway undertakings in that country.* Vol.XLVIII, 32 pp.

(40) Pérez, S. (2018). "Railroads and the Rural to Urban Transition: Evidence from 19th-century Argentina." Working Paper, University of California, Davis.

(41) Sanyal, Nalinaksha. (1930). *Development of Indian Railways.* University of Calcutta, Calcutta.

(42) Sharp, H. (1918). *Progress of Education in India: Quinquennial Review. 1912-1917.* Calcutta: Government printing.

(43) Summerhill, W. R. (2005). "Big Social Savings in a Small Laggard Economy: Railroad-Led Growth in Brazil." *Journal of Economic History* 65(1): 72-102.

(44) Tang, John. (2017). "The Engine and the Reaper: Industrialization and Mortality in Late Nineteenth Century Japan." *Journal of Health Economics*, 56 (December): 145-162.

(45) Thorner, Daniel. (1951). "Great Britain and the Development of India's Railways." *Journal of Economic History* 11(4): 389-402.

(46) Thorner, Daniel. (1955). "The Pattern of Railway Development in India." *Far Eastern Quarterly*, XIV: 201-6.

(47) Walby, Kevin and Michael Haan. (2012). "Caste Confusion and Census Enumeration in Colonial India, 1871–1921." *Histoire sociale/Social history*, 45 (no. 90): 301-318.

(48) Willmott, C. J. and K. Matsuura. (2018). "Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1900 - 2017)." `http://climate.geog.udel.edu/~climate/html_pages/download.html#T2017`.

8. Tables

Table 1. Summary Statistics: Cohort

	Under 10	10-15	15-20	20 & Above
1901				
Total	0.77%	4.76%	6.63%	6.45%
Male	1.30%	7.82%	11.68%	12.04%
Female	0.25%	1.02%	1.21%	0.69%
English	0.08%	0.50%	0.84%	0.63%
English Male	0.11%	0.77%	1.43%	1.07%
English Female	0.05%	0.16%	0.20%	0.13%
Non-English	0.69%	4.25%	5.79%	5.82%
Non-English Male	1.19%	7.05%	10.25%	10.97%
Non-English Female	0.19%	0.85%	1.01%	0.56%
Cohorts Years of Railroad Exposure	21.51	18.13	14.31	10.83
1911				
Total	0.78%	5.67%	7.63%	7.26%
Male	1.21%	8.96%	13.04%	13.09%
Female	0.33%	1.60%	1.83%	1.07%
English	0.12%	0.70%	1.20%	0.90%
English Male	0.13%	1.00%	1.99%	1.48%
English Female	0.10%	0.30%	0.36%	0.23%
Non-English	0.66%	4.97%	6.43%	6.36%
Non-English Male	1.08%	7.96%	11.04%	11.61%
Non-English Female	0.24%	1.30%	1.47%	0.83%
Cohorts Years of Railroad Exposure	30.59	26.88	22.38	18.13
1921				
Total	1.02%	6.67%	9.42%	8.41%
Male	1.53%	10.08%	15.38%	14.77%
Female	0.51%	2.43%	2.87%	1.62%
English	0.14%	0.98%	2.07%	1.23%
English Male	0.17%	1.41%	3.25%	2.03%
English Female	0.10%	0.43%	0.80%	0.32%
Non-English	0.89%	5.69%	7.35%	7.18%
Non-English Male	1.36%	8.67%	12.13%	12.74%
Non-English Female	0.41%	2.00%	2.07%	1.31%
Cohorts Years of Railroad Exposure	40.12	36.28	31.53	26.88

Note: See text for details.

Table 2. Summary Statistics by Cross Section

	Mean	SD	Min	Max	N
Literacy 1881	3.15%	2.43%	0.27%	17.66%	198
Literacy 1891	4.18%	3.55%	0.60%	35.23%	199
Literacy 1901	4.77%	3.33%	0.69%	24.82%	203
Literacy 1911	5.35%	3.95%	0.86%	32.13%	203
Literacy 1921	6.25%	4.62%	1.27%	41.88%	203
Male Litearcy 1881	5.83%	3.99%	0.52%	30.52%	198
Male Litearcy 1891	7.55%	4.80%	1.10%	35.23%	199
Male Litearcy 1901	8.67%	5.26%	1.34%	35.99%	203
Male Litearcy 1911	9.42%	5.83%	1.65%	42.13%	203
Male Litearcy 1921	10.65%	6.53%	2.29%	50.15%	203
Female Litearcy 1881	0.27%	0.72%	0.01%	6.33%	197
Female Litearcy 1891	0.40%	0.99%	0.04%	8.73%	198
Female Litearcy 1901	0.65%	1.41%	0.02%	11.49%	203
Female Litearcy 1911	0.99%	1.97%	0.05%	16.45%	203
Female Litearcy 1921	1.50%	2.62%	0.12%	24.30%	203
English Litearcy 1901	0.50%	1.11%	0.00%	10.31%	203
English Litearcy 1911	0.72%	1.62%	0.01%	14.20%	203
English Litearcy 1921	0.99%	1.97%	0.04%	19.15%	203
Railroad Years 1881	7.58	9.18	0	28	198
Railroad Years 1891	13.86	12.62	0	38	199
Railroad Years 1901	21.51	15.18	0	48	203
Railroad Years 1911	30.59	16.65	0	58	203
Railroad Years 1921	40.12	17.62	0	68	203
Railroad Indicator 1881	52.02%	50.09%	0	1	198
Railroad Indicator 1891	73.37%	44.32%	0	1	199
Railroad Indicator 1901	87.19%	33.50%	0	1	203
Railroad Indicator 1911	93.60%	24.54%	0	1	203
Railroad Indicator 1921	96.06%	19.50%	0	1	203

Note: See text for details. We do not have English literacy before 1901.

TABLE 3. Synthetic Panel: Cohort, District and Year Fixed Effects

	(1)	(2)	(3)
	Total	Male	Female
	Literacy		
Cohort Years of Railroad Exposure	0.0202*** (0.0070)	0.0224*** (0.0071)	0.0079 (0.0078)
Obs.	1,609	1,609	1,608
	English Literacy		
Cohort Years of Railroad Exposure	0.0234*** (0.0078)	0.0266*** (0.0086)	0.0050 (0.0079)
Obs.	1,598	1,597	1,536
	Non-English Literacy		
Cohort Years of Railroad Exposure	0.0212*** (0.0074)	0.0235*** (0.0075)	0.0080 (0.0081)
Obs.	1,607	1,607	1,606
Years	1911-1921	1911-1921	1911-1921

Note: Robust standard errors clustered at district level in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$ The unit of analysis is log literacy at the cohort-year level. All the regressions include fixed effects for district, cohort $\times$ province and year$\times$ province.

Table 4. Cross-Section: Ordinary Least Squares

	(1)	(2) Literacy	(3)	(4) Male	(5) Female	(6) English
			Year 1881			
Railroad Years	0.0269***	0.0191***	0.0131***	0.0122***	0.0373***	
	(0.0047)	(0.0036)	(0.0025)	(0.0024)	(0.0063)	
Obs	198	198	198	198	197	
			Year 1891			
Railroad Years	0.0140***	0.0119***	0.0096***	0.0084***	0.0232***	
	(0.0034)	(0.0026)	(0.0020)	(0.0018)	(0.0042)	
Obs	199	199	199	199	198	
			Year 1901			
Railroad Years	0.0105***	0.0105***	0.0078***	0.0068***	0.0200***	0.0236***
	(0.0029)	(0.0028)	(0.0018)	(0.0017)	(0.0037)	(0.0038)
			Year 1911			
Railroad Years	0.0094***	0.0101***	0.0070***	0.0060***	0.0151***	0.0213***
	(0.0026)	(0.0023)	(0.0016)	(0.0015)	(0.0030)	(0.0032)
Obs	203	203	203	203	203	203
			Year 1921			
Railroad Years	0.0073***	0.0090***	0.0058***	0.0045***	0.0134***	0.0164***
	(0.0025)	(0.0022)	(0.0016)	(0.0015)	(0.0029)	(0.0031)
Obs	203	203	203	203	203	203
Controls	No	No	Yes	Yes	Yes	Yes
FE	No	Province	Province	Province	Province	Province

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The controls include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

TABLE 5. Cross-Section: Instrument 1852 Kennedy Plan

	(1)	(2)	(3)	(4)	(5)	(6)
		Literacy		Male	Female	English
			Year 1881			
Railroad Years	0.0441***	0.0242***	0.0085	0.0076	0.0321**	
	(0.0103)	(0.0073)	(0.0061)	(0.0060)	(0.0155)	
Obs	198	198	198	198	197	
KPF	41.87	37.76	24.46	24.46	25.06	
			Year 1891			
Railroad Years	0.0324***	0.0146**	0.0067	0.0053	0.0262**	
	(0.0085)	(0.0068)	(0.0052)	(0.0051)	(0.0110)	
Obs	199	199	199	199	198	
KPF	30.51	29.53	19.11	19.11	19.05	
			Year 1901			
Railroad Years	0.0259***	0.0117*	0.0090	0.0076	0.0231*	0.0105
	(0.0086)	(0.0066)	(0.0058)	(0.0058)	(0.0128)	(0.0143)
Obs	203	203	203	203	203	203
KPF	21.51	22.49	13.50	13.50	13.50	13.50
			Year 1911			
Railroad Years	0.0300***	0.0132**	0.0078	0.0059	0.0193*	0.0291**
	(0.0087)	(0.0061)	(0.0057)	(0.0054)	(0.0111)	(0.0133)
Obs	203	203	203	203	203	203
KPF	19.80	19.91	10.99	10.99	10.99	10.99
			Year 1921			
Railroad Years	0.0293***	0.0122**	0.0148**	0.0111*	0.0389***	0.0322**
	(0.0089)	(0.0060)	(0.0061)	(0.0056)	(0.0127)	(0.0135)
Obs	203	203	203	203	203	203
KPF	18.06	17.92	11.22	11.22	11.22	11.22
Controls	No	No	Yes	Yes	Yes	Yes
FE	No	Province	Province	Province	Province	Province

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The controls include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table 6. Kennedy, Military Cantonment and Pre-1850 Cities Instruments

	(1)	(2)	(3)	(4)	(5)	(6)
		Literacy		Male	Female	English
			Year 1881			
Railroad Years	0.0388***	0.0237***	0.0165***	0.0160***	0.0482***	
	(0.0072)	(0.0053)	(0.0042)	(0.0040)	(0.0109)	
KPF	70.52	50.10	35.60	35.60	35.61	
P-value Over-id Test	0.167	0.986	0.172	0.138	0.317	
			Year 1891			
Railroad Years	0.0233***	0.0183***	0.0164***	0.0145***	0.0360***	
	(0.0049)	(0.0041)	(0.0040)	(0.0037)	(0.0072)	
KPF	68.76	44.02	31.16	31.16	30.61	
P-value Over-id Test	0.0918	0.805	0.0968	0.107	0.539	
			Year 1901			
Railroad Years	0.0187***	0.0173***	0.0149***	0.0144***	0.0316***	0.0453***
	(0.0044)	(0.0043)	(0.0037)	(0.0038)	(0.0069)	(0.0080)
KPF	56.64	35.91	21.82	21.82	21.82	21.82
P-value Over-id Test	0.523	0.408	0.374	0.307	0.318	0.0209
			Year 1911			
Railroad Years	0.0189***	0.0182***	0.0149***	0.0139***	0.0278***	0.0423***
	(0.0042)	(0.0040)	(0.0036)	(0.0035)	(0.0063)	(0.0077)
KPF	50.02	31.08	18.66	18.66	18.66	18.66
P-value Over-id Test	0.168	0.557	0.213	0.156	0.479	0.561
			Year 1921			
Railroad Years	0.0163***	0.0172***	0.0145***	0.0126***	0.0302***	0.0368***
	(0.0042)	(0.0041)	(0.0037)	(0.0035)	(0.0065)	(0.0076)
KPF	45.20	27.96	17.06	17.06	17.06	17.06
P-value Over-id Test	0.102	0.444	0.958	0.857	0.569	0.870
Controls	No	No	Yes	Yes	Yes	Yes
FE	No	Province	Province	Province	Province	Province

Note: Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$ The controls include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table 7. Enrolment

	(1)	(2)	(3)	(4)	(5)	(6)
	Enrolment	Enrolment	Primary Enrolment	Primary Enrolment	Secondary Enrolment	Secondary Enrolment
Panel: District and Year Fixed Effects						
Railroad Years	-0.0041	0.0022	-0.0201	-0.0184	0.0277*	0.0302**
	(0.0177)	(0.0107)	(0.0202)	(0.0156)	(0.0164)	(0.0117)
Obs	919	641	923	645	919	641
Year	All	1894/1897 1901/1905 /1911	All	1894/1897 1901/1905 /1911	All	1894/1897 1901/1905 /1911
Cross-Section						
1901						
Railroad Years	0.0050**	0.0170**	0.0031	0.0104	0.0159***	0.0475***
	(0.0021)	(0.0071)	(0.0024)	(0.0083)	(0.0030)	(0.0126)
Obs	179	179	182	182	179	179
Model	OLS	IV	OLS	IV	OLS	IV
1911						
Railroad Years	0.0035**	0.0074	0.0013	0.0037	0.0140***	0.0346***
	(0.0017)	(0.0069)	(0.0018)	(0.0074)	(0.0030)	(0.0133)
Obs	177	177	178	178	177	177
Model	OLS	IV	OLS	IV	OLS	IV

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The cross-sectional models include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table 8. Mediators: Total Literacy, OLS

	(1)	(2)	(3)	(4)	(5)	(6)
	Ag Income		Land Taxes Per-Capita		Income Taxes Per-Capita	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0057***	0.0065***	0.0065***	0.0069***	0.0035**	0.0031**
	(0.0017)	(0.0015)	(0.0017)	(0.0014)	(0.0017)	(0.0014)
Ln(Ag. Income)	-0.0436	-0.0565				
	(0.0516)	(0.0471)				
Ln(Land Taxes Per-Capita)			-0.0080	0.0402		
			(0.0467)	(0.0441)		
Ln(Income Taxes Per-Capita)					0.1834***	0.1864***
					(0.0350)	(0.0292)
% of Total Effect Mediated	-0.064	-0.04	-0.006	0.011	0.464	0.506
	Share Workers					
	Industry		Services		Public Adm	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0065***	0.0064***	0.0049***	0.0052***	0.0058***	0.0041***
	(0.0018)	(0.0015)	(0.0017)	(0.0014)	(0.0017)	(0.0015)
Ln(Share Workers, Industry)	0.0898	0.1830***				
	(0.0692)	(0.0633)				
Ln(Share Workers, Services)			0.3635***	0.3832***		
			(0.0770)	(0.0623)		
Ln(Share Workers, Public Adm)					0.2047***	0.3007***
					(0.0629)	(0.0501)
% of Total Effect Mediated	0.008	0.077	0.241	0.253	0.116	0.411

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The outcome is log literacy in the respective year. These cross-sectional models include province fixed effects; social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table 9. Mediators: Secondary Enrolment, OLS

	(1)	(2)	(3)	(4)	(5)	(6)
	Ag Income		Land Taxes Per-Capita		Income Taxes Per-Capita	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0187***	0.0131***	0.0154***	0.0138***	0.0079***	0.0095***
	(0.0032)	(0.0032)	(0.0032)	(0.0032)	(0.0028)	(0.0031)
Ln(Ag. Income)	-0.1561	0.0778				
	(0.1000)	(0.1156)				
Ln(Land Taxes Per-Capita)			0.1135	0.0661		
			(0.0859)	(0.0986)		
Ln(Income Taxes Per-Capita)					0.5266***	0.2641***
					(0.0631)	(0.0646)
%Pct of Total Effect Mediated	-0.059	0.008	0.0312	0.0118	0.502	0.311

	Share Workers					
	Industry		Services		Public Adm	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0161***	0.0139***	0.0132***	0.0126***	0.0152***	0.0118***
	(0.0032)	(0.0032)	(0.0031)	(0.0031)	(0.0032)	(0.0033)
Ln(Share Workers, Industry)	0.3636***	0.1861				
	(0.1273)	(0.1353)				
Ln(Share Workers, Services)			0.6920***	0.4650***		
			(0.1429)	(0.1516)		
Ln(Share Workers, Public Adm)					0.3448***	0.3211**
					(0.1182)	(0.1283)
% of Total Effect Mediated	0.028	0.045	0.202	0.138	0.0785	0.188

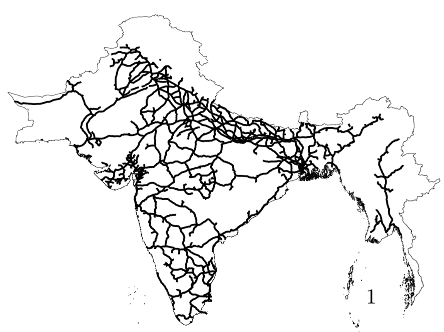
Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The outcome is log secondary enrolment in the respective year. These cross-sectional models include province fixed effects; social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

9. Figures

Figure 1. Rail Network 1881-1921

(a) 1881

(b) 1901



(c) 1921

Figure 2. Distribution of Total Literacy

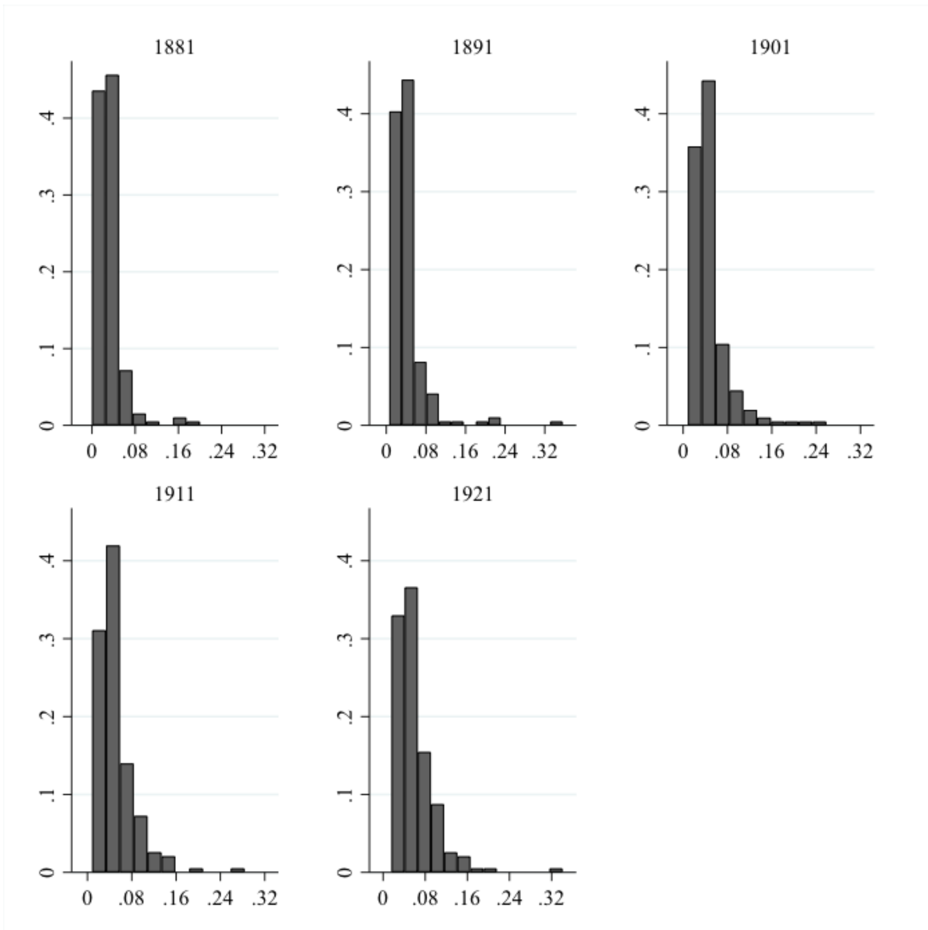


Figure 3. Distribution of Male Literacy

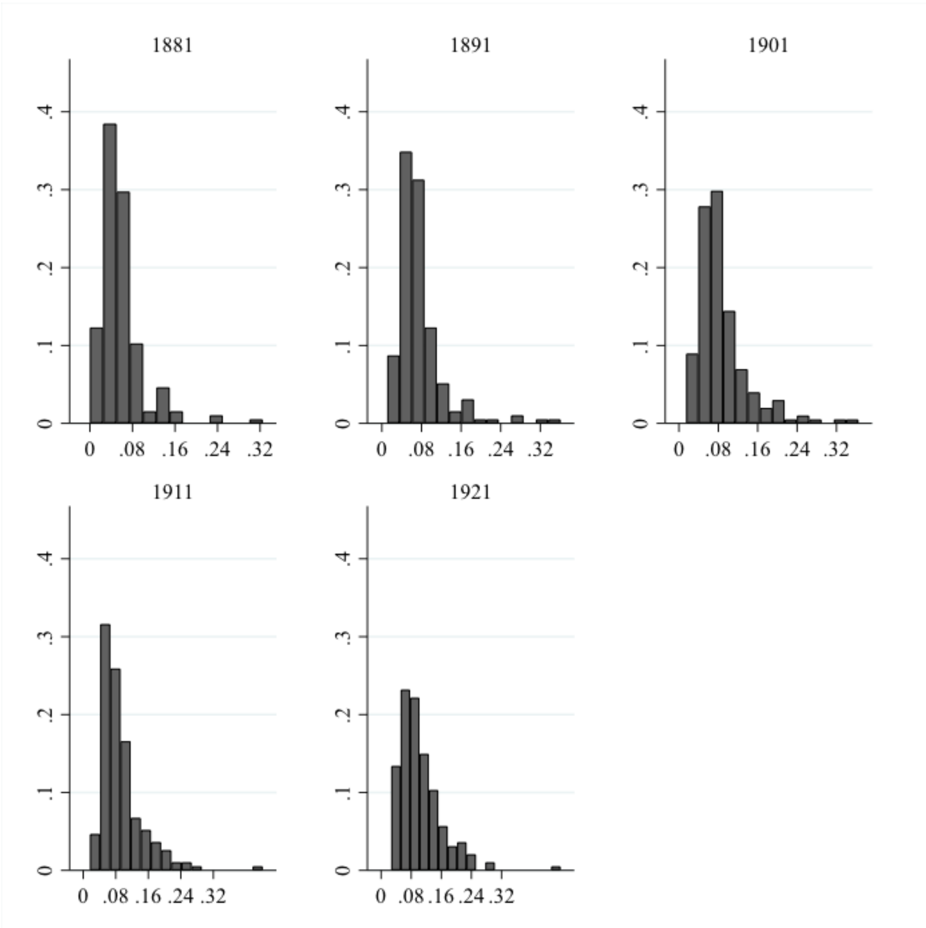


Figure 4. Distribution of Female Literacy

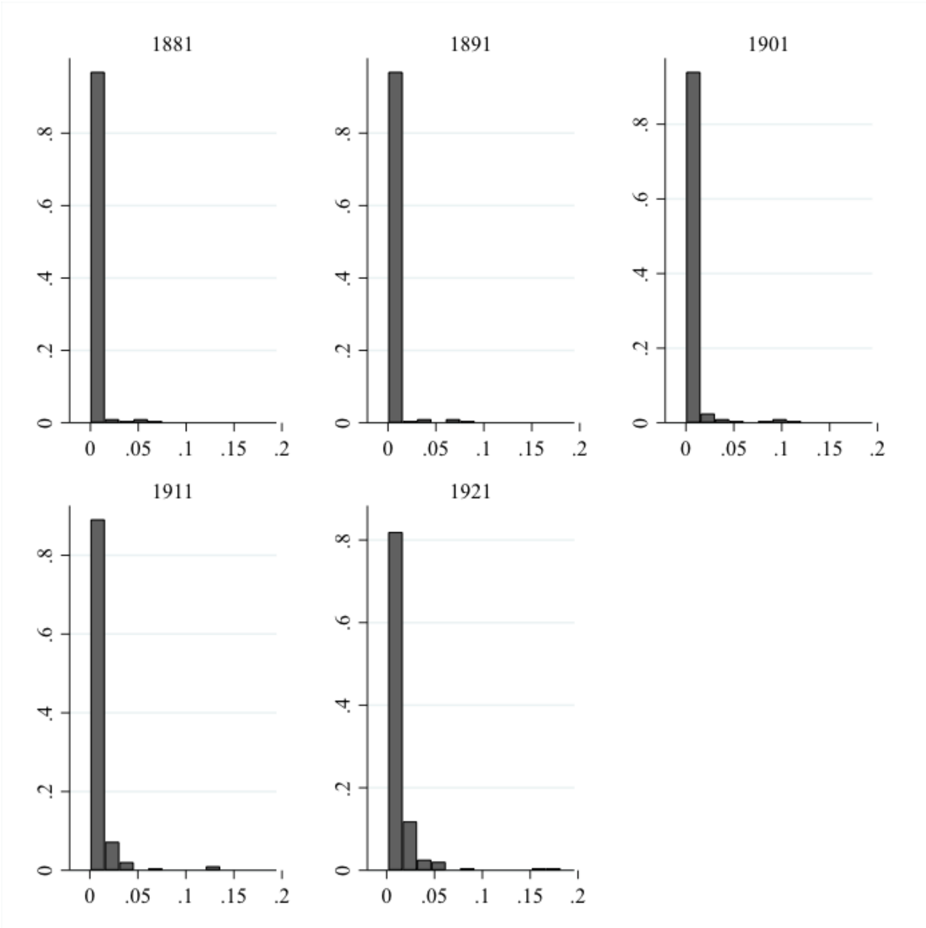
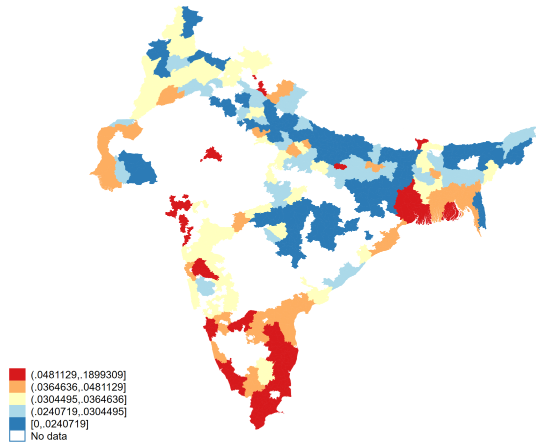
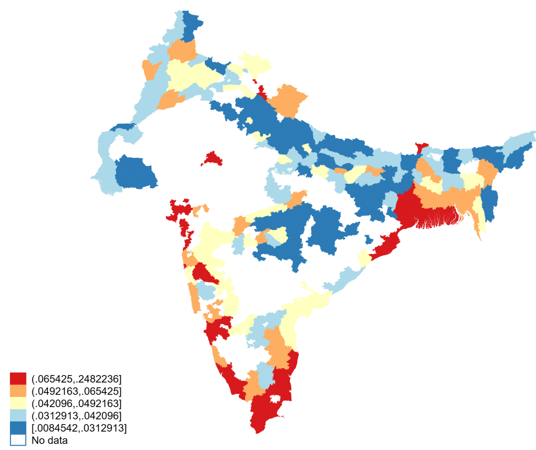


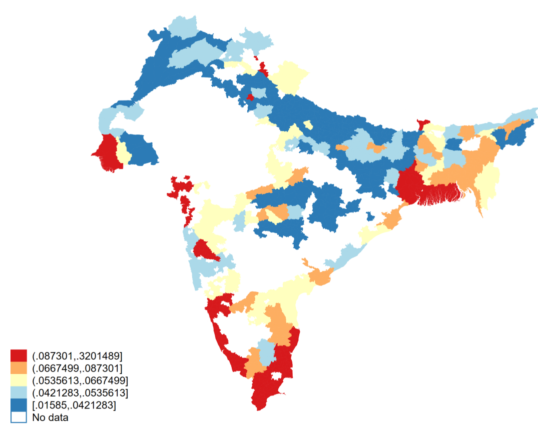
Figure 5. Map of Total Literacy, 1881-1921, Quintiles



(a) 1881



(b) 1901



(c) 1921

Figure 6. Scatterplot of Railroad Years and Literacy, 1881

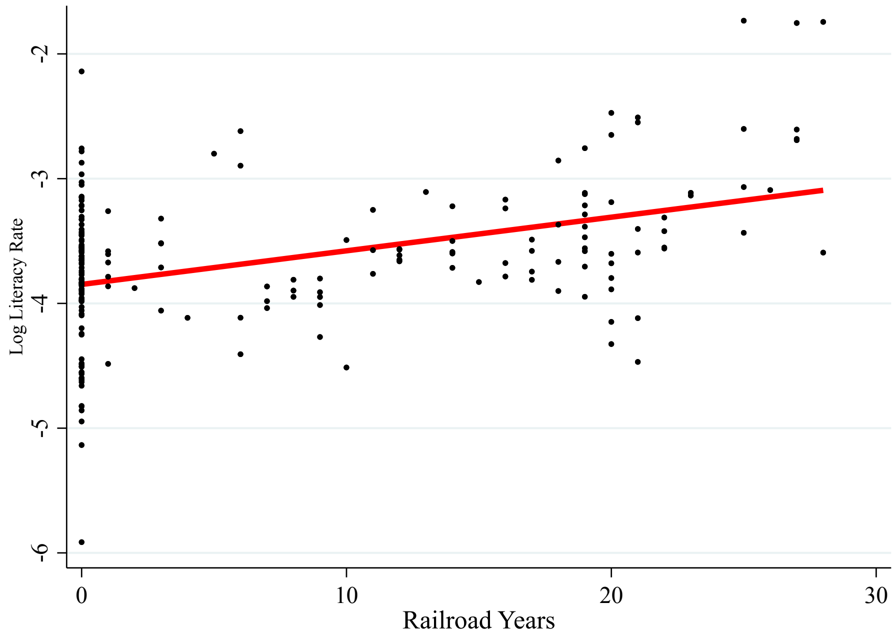


Figure 7. Scatterplot of Railroad Years and Literacy, 1921

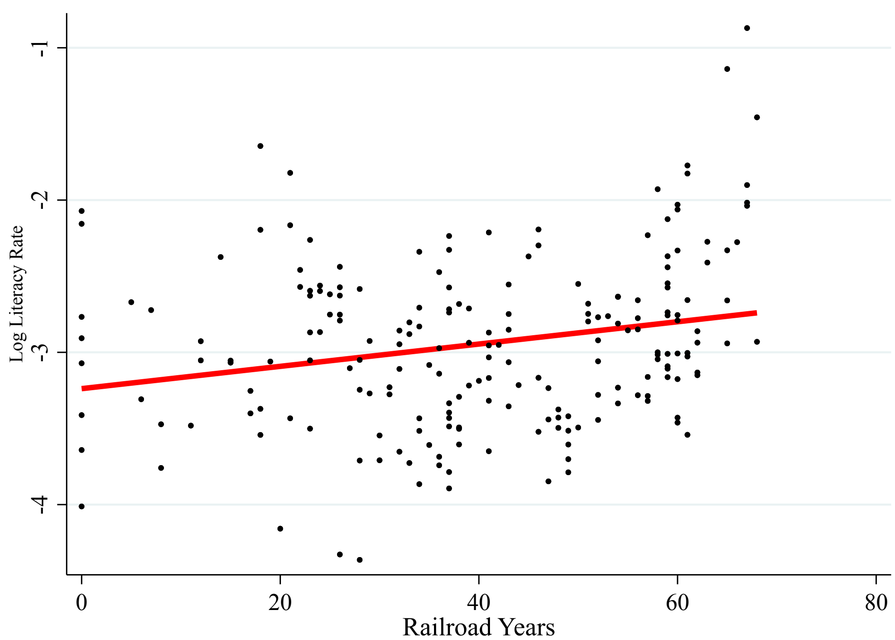


Figure 8. Map of 1852 Kennedy Plan

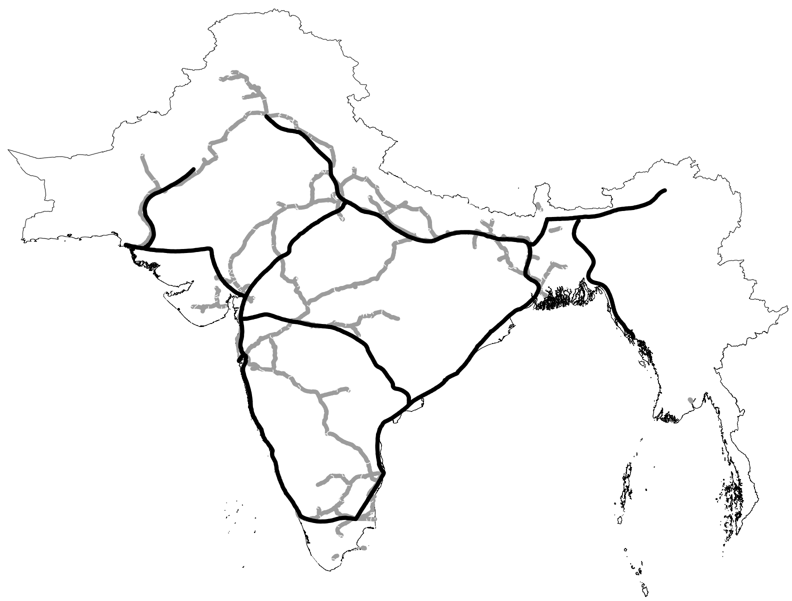
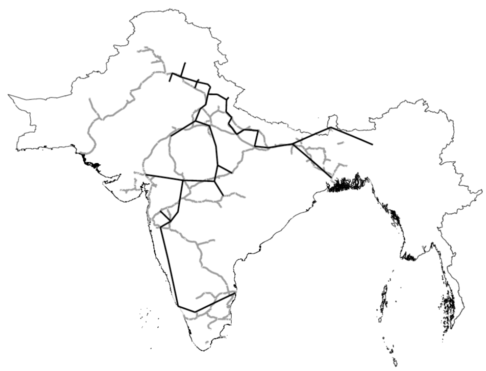


FIGURE 9. Map of Spanning Trees



(a) Tree Spanning Military Cantons, 1864

(b) Tree Spanning Cities, 1850

Appendix A. Appendix Tables

Table A1. Synthetic Panel: Conley Standard Errors, Cutoff 200 km

	(1)	(2)	(3)
	Total	Male	Female
	Literacy		
Cohort Years	0.0202***	0.0224***	0.0079
of Railroad Exposure	(0.0070)	(0.0071)	(0.0078)
Obs.	1,609	1,609	1,608
	English Literacy		
Cohort Years	0.0234***	0.0266***	0.0050
of Railroad Exposure	(0.0078)	(0.0086)	(0.0079)
Obs.	1,598	1,597	1,536
	Non-English Literacy		
Cohort Years	0.0212***	0.0235***	0.0080
of Railroad Exposure	(0.0074)	(0.0075)	(0.0081)
Obs.	1,607	1,607	1,606
Years	1911-1921	1911-1921	1911-1921

Note: Conley standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The unit of analysis is log literacy at the cohort*year level. All the regressions include fixed effects for district, cohort× province and year× province.

Table A2. Synthetic Panel: Dropping Bombay, Calcutta, Delhi and Madras

	(1)	(2)	(3)
	Total	Male	Female
	Literacy		
Cohort Years	0.0191***	0.0212***	0.0075
of Railroad Exposure	(0.0071)	(0.0072)	(0.0079)
Obs.	1,577	1,577	1,576
	English Literacy		
Cohort Years	0.0221***	0.0252***	0.0044
of Railroad Exposure	(0.0076)	(0.0085)	(0.0078)
Obs.	1,566	1,565	1,504
	Non-English Literacy		
Cohort Years	0.0200***	0.0222***	0.0076
of Railroad Exposure	(0.0075)	(0.0076)	(0.0082)
Obs.	1,576	1,576	1,575
Years	1911-1921	1911-1921	1911-1921

Note: Robust standard errors clustered at district level in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The unit of analysis is log literacy at the cohort*year level. All the regressions include fixed effects for district, cohort× province and year× province.

Table A3. Synthetic Panel: Dropping Provinces, Total Literacy

	(1)	(2)	(3)	(4)
Cohort Years of Railroad Exposure	0.0202*** (0.0070)	0.0147** (0.0071)	0.0176** (0.0072)	0.0196** (0.0077)
Drop Province	Ajmer	Assam	Bengal	Bihar & Orissa
Cohort Years of Railroad Exposure	0.0204*** (0.0078)	0.0216*** (0.0080)	0.0202*** (0.0070)	0.0205*** (0.0072)
Drop Province	Bombay	CP	Coorg	Madras
Cohort Years of Railroad Exposure	0.0207*** (0.0075)	0.0260*** (0.0071)	0.0204*** (0.0071)	
Drop Province	NWFP	Punjab	UP	

Note: Robust standard errors clustered at district level in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The unit of analysis is log literacy at the cohort*year level. All the regressions include fixed effects for district, cohort× province and year× province.

Table A4. Synthetic Panel: Excluding Cohort 20 and Above

	(1)	(2)	(3)
	Total	Male	Female
	Literacy		
Cohort Years	0.0235***	0.0271***	0.0131
of Railroad Exposure	(0.0090)	(0.0093)	(0.0109)
Obs.	1,206	1,206	1,205
	English Literacy		
Cohort Years	0.0262	0.0305*	-0.0027
of Railroad Exposure	(0.0160)	(0.0164)	(0.0131)
Obs.	1,195	1,194	1,133
	Non-English Literacy		
Cohort Years	0.0246**	0.0285***	0.0128
of Railroad Exposure	(0.0096)	(0.0099)	(0.0109)
Obs.	1,204	1,204	1,203
Years	1911-1921	1911-1921	1911-1921
Years	1911-1921	1911-1921	1911-1921

Note: Robust standard errors clustered at district level in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The unit of analysis is log literacy at the cohort*year level. All the regressions include fixed effects for district, cohort $\times$ province and year$\times$ province.

Table A5. Synthetic Panel: Alternative Age Band for Cohort Years

	(1)	(2)	(3)
	Total	Male	Female
	Literacy		
Cohort Years	0.0186**	0.0206**	0.0028
of Railroad Exposure	(0.0087)	(0.0085)	(0.0118)
Obs.	1,609	1,609	1,608
	English Literacy		
Cohort Years	0.0211**	0.0231**	0.0054
of Railroad Exposure	(0.0107)	(0.0115)	(0.0162)
Obs.	1,598	1,597	1,536
	Non-English Literacy		
Cohort Years	0.0187**	0.0207**	0.0030
of Railroad Exposure	(0.0088)	(0.0087)	(0.0125)
Obs.	1,607	1,607	1,606
Years	1911-1921	1911-1921	1911-1921

Note: Robust standard errors clustered at district level in parentheses.
*** p<0.01, ** p<0.05, * p<0.1 The unit of analysis is log literacy at the cohort*year level. All the regressions include fixed effects for district, cohort × province and year× province.

Table A6. Synthetic Cohort: 1901 Census Year

	(1)	(2)	(3)
	Total	Male	Female
	Literacy		
Cohort Years	0.0066	0.0081	-0.0005
of Railroad Exposure	(0.0053)	(0.0052)	(0.0057)
Obs.	812	812	811
	English Literacy		
Cohort Years	0.0344***	0.0293***	-0.0095
of Railroad Exposure	(0.0103)	(0.0102)	(0.0104)
Obs.	809	805	746
	Non-English Literacy		
Cohort Years	0.0061	0.0070	0.0033
of Railroad Exposure	(0.0053)	(0.0055)	(0.0058)
Obs.	812	812	811
Years	1901	1901	1901

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The unit of analysis is log literacy for the 1901 cohort. All the regressions include district and cohort × province fixed effects.

Table A7. Cross-Section: OLS with Conley SE

	(1)	(2)	(3)	(4)	(5)	(6)
		Literacy		Male	Female	English
			Year 1881			
Railroad Years	0.0269***	0.0191***	0.0131***	0.0122***	0.0373***	
	(0.0051)	(0.0037)	(0.0020)	(0.0019)	(0.0045)	
Obs	198	198	198	198	197	
			Year 1891			
Railroad Years	0.0140***	0.0119***	0.0096***	0.0084***	0.0232***	
	(0.0037)	(0.0033)	(0.0016)	(0.0016)	(0.0028)	
Obs	199	199	199	199	198	
			Year 1901			
Railroad Years	0.0105**	0.0105***	0.0078***	0.0068***	0.0200***	0.0236***
	(0.0044)	(0.0040)	(0.0021)	(0.0020)	(0.0037)	(0.0032)
Obs	203	203	203	203	203	203
			Year 1911			
Railroad Years	0.0094***	0.0101***	0.0070***	0.0060***	0.0151***	0.0213***
	(0.0034)	(0.0029)	(0.0015)	(0.0013)	(0.0028)	(0.0026)
Obs	203	203	203	203	203	203
			Year 1921			
Railroad Years	0.0073**	0.0090***	0.0058***	0.0045***	0.0134***	0.0164***
	(0.0031)	(0.0022)	(0.0015)	(0.0014)	(0.0026)	(0.0025)
Obs	203	203	203	203	203	203
Controls	No	No	Yes	Yes	Yes	Yes
FE	No	Province	Province	Province	Province	Province

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The controls include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A8. Cross-Section: Grid Cell Fixed Effects

	(1) Total	(2) Male	(3) Female	(4) English
Year 1881				
Railroad Years	0.0110***	0.0105***	0.0341***	
	(0.0034)	(0.0032)	(0.0090)	
Obs	198	198	197	
Year 1891				
Railroad Years	0.0077***	0.0066***	0.0255***	
	(0.0023)	(0.0021)	(0.0051)	
Obs	199	199	198	
Year 1901				
Railroad Years	0.0051**	0.0042**	0.0190***	0.0257***
	(0.0021)	(0.0020)	(0.0049)	(0.0055)
Obs	203	203	203	203
Year 1911				
Railroad Years	0.0060***	0.0048***	0.0175***	0.0249***
	(0.0019)	(0.0017)	(0.0036)	(0.0044)
Obs	203	203	203	203
Year 1921				
Railroad Years	0.0059***	0.0043**	0.0172***	0.0218***
	(0.0020)	(0.0018)	(0.0036)	(0.0043)
Obs	203	203	203	203
Controls	Yes	Yes	Yes	Yes
FE	Grid Cell	Grid Cell	Grid Cell	Grid Cell

Note: Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$ The controls include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A9. Matching ATT Estimates

	(1)	(2)	(3)
	Total	Male	Female
	1881		
Railroads Indicator	0.215***	0.201***	0.505***
	(0.054)	(0.057)	(0.112)
	1891		
Railroads Indicator	0.212***	0.181***	0.503***
	(0.047)	(0.048)	(0.114)

Note: Abadie Imbens standard errors in paranthese. *** p<0.01, ** p<0.05, * p<0.1 We match districts using the following controls: social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A10. Cross-Section: OLS with Market Access

	(1)	(2) Literacy	(3)	(4) Male	(5) Female	(6) English
			Year 1881			
Market Access	0.0058**	0.0067***	0.0064***	0.0060***	0.0123	
	(0.0023)	(0.0020)	(0.0018)	(0.0016)	(0.0088)	
Obs	198	198	198	198	197	
			Year 1891			
Market Access	0.0114*	0.0108**	0.0073***	0.0064***	0.0091*	
	(0.0067)	(0.0048)	(0.0017)	(0.0011)	(0.0050)	
Obs	199	199	199	199	198	
			Year 1901			
Market Access	0.0098**	0.0087***	0.0041***	0.0031**	0.0099**	0.0165***
	(0.0043)	(0.0026)	(0.0015)	(0.0013)	(0.0041)	(0.0049)
Obs	200	200	200	200	200	200
			Year 1911			
Market Access	0.0113***	0.0100***	0.0058***	0.0043***	0.0122***	0.0173***
	(0.0039)	(0.0023)	(0.0016)	(0.0014)	(0.0034)	(0.0042)
Obs	200	200	200	200	200	200
			Year 1921			
Market Access	0.0100***	0.0087***	0.0052***	0.0037***	0.0113***	0.0138***
	(0.0024)	(0.0014)	(0.0014)	(0.0011)	(0.0031)	(0.0032)
Obs	200	200	200	200	200	200
Controls	No	No	Yes	Yes	Yes	Yes
FE	No	Province	Province	Province	Province	Province

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The controls include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

TABLE A11. First Stage: Instrument 1852 Kennedy Plan Instrument

	(1) 1881	(2) 1891	(3) 1901	(4) 1911	(5) 1921
Ln (Distance from Kennedy Plan)	-0.6938***	-0.8189***	-0.7834***	-0.7841**	-0.8190***
	(0.1403)	(0.1873)	(0.2132)	(0.2365)	(0.2445)
GIS controls	Yes	Yes	Yes	Yes	Yes
Religious controls	Yes	Yes	Yes	Yes	Yes
Province FE	Yes	Yes	Yes	Yes	Yes
KFP	24.46	19.11	13.50	10.99	11.22
Obs	198	199	203	203	203

Note: Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$ The dependent variable in the first stage is railroad years. The controls include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A12. Cross-Section: Military and Chandler Tree IV

	(1)	(2)	(3)	(4)	(5)	(6)
	Total		Male		Female	
	Military IV	Chandler IV	Military IV	Chandler IV	Military IV	Chandler IV
Year 1881						
Railroad Years	0.0223***	0.0200***	0.0224***	0.0195***	0.0567***	0.0588***
	(0.0064)	(0.0060)	(0.0062)	(0.0059)	(0.0162)	(0.0147)
Obs	198	198	198	198	197	197
KPF	22.88	23.91	22.88	23.91	22.36	24.46
Year 1891						
Railroad Years	0.0177***	0.0233***	0.0168***	0.0197***	0.0391***	0.0407***
	(0.0049)	(0.0067)	(0.0047)	(0.0058)	(0.0097)	(0.0100)
Obs	199	199	199	199	198	198
KPF	26	23.87	26	23.87	25.86	24.52
Year 1901						
Railroad Years	0.0188***	0.0144***	0.0185***	0.0141***	0.0405***	0.0270***
	(0.0051)	(0.0049)	(0.0051)	(0.0049)	(0.0096)	(0.0093)
Obs	203	203	203	203	203	203
KPF	24.41	19.08	24.41	19.08	24.41	19.08
Year 1911						
Railroad Years	0.0151***	0.0201***	0.0144***	0.0192***	0.0280***	0.0341***
	(0.0045)	(0.0056)	(0.0043)	(0.0054)	(0.0080)	(0.0088)
Obs	203	203	203	203	203	203
KPF	23.95	14.98	23.95	14.98	23.95	14.98
Year 1921						
Railroad Years	0.0151***	0.0135***	0.0140***	0.0118**	0.0247***	0.0311***
	(0.0047)	(0.0052)	(0.0044)	(0.0049)	(0.0083)	(0.0094)
Obs	203	203	203	203	203	203
KPF	22.55	12.25	22.55	12.25	22.55	12.25
Controls	Yes	Yes	Yes	Yes	Yes	Yes
FE	Province	Province	Province	Province	Province	Province

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The controls include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A13. Mediator: Agricultural Income, Instrumental Variables

	(1)	(2)	(3)	(4)	(5)	(6)
	Enrolment		Primary Enrolment		Secondary Enrolment	
	1901	1911	1901	1911	1901	1911
Total Effect	0.0121**	0.0088	0.0108	0.0054	0.0293***	0.0313***
(Railroads)	(0.0057)	(0.0060)	(0.0070)	(0.0060)	(0.0092)	(0.0120)
Direct Effect	0.0117	0.0046**	0.0100	0.0022	0.0293**	0.0174***
(Unmediated)	(0.0082)	(0.0022)	(0.0097)	(0.0021)	(0.0124)	(0.0059)
Indirect Effect	-0.0001	0.0042	0.0000	0.0032	-0.0001	0.0140
(Mediated)	(0.0063)	(0.0068)	(0.0073)	(0.0059)	(0.0105)	(0.0200)
Obs.	149	145	149	145	149	145
Controls	All	All	All	All	All	All
FE	Province	Province	Province	Province	Province	Province
	Literacy		Male Literacy		English Literacy	
	1901	1911	1901	1911	1901	1911
Total Effect	0.0218***	0.0203***	0.0216***	0.0196***	0.0301***	0.0265***
(Railroads)	(0.0069)	(0.0054)	(0.0069)	(0.0053)	(0.0114)	(0.0088)
Direct Effect	0.0346	0.0419	0.0339	0.0407	0.0425	0.0431
(Unmediated)	(0.0439)	(0.1477)	(0.0442)	(0.1462)	(0.0397)	(0.1014)
Indirect Effect	-0.0132	-0.0205	-0.0132	-0.0203	-0.0106	-0.0137
(Mediated)	(0.0330)	(0.1088)	(0.0331)	(0.1076)	(0.0278)	(0.0740)
Obs.	163	157	163	157	163	157
Controls	All	All	All	All	All	All
FE	Province	Province	Province	Province	Province	Province

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The cross-sectional models include social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A14. Mediators: Male Literacy, OLS

	(1)	(2)	(3)	(4)	(5)	(6)
	Ag Income		Land Taxes Per-Capita		Income Taxes Per-Capita	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0048***	0.0053***	0.0057***	0.0061***	0.0032*	0.0027*
	(0.0016)	(0.0014)	(0.0017)	(0.0014)	(0.0017)	(0.0014)
Ln(Ag. Income)	-0.0323	-0.0428				
	(0.0503)	(0.0453)				
Ln(Land Taxes Per-Capita)			-0.0073	0.0382		
			(0.0461)	(0.0423)		
Ln(Income Taxes Per-Capita)					0.1460***	0.1589***
					(0.0351)	(0.0286)
% of Total Effect Mediated	-0.056	-0.0361	-0.006	0.012	0.428	0.498

	Share Workers					
	Industry		Services		Public Adm	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0055***	0.0054***	0.0041**	0.0044***	0.0049***	0.0035**
	(0.0017)	(0.0014)	(0.0017)	(0.0014)	(0.0017)	(0.0014)
Ln(Share Workers, Industry)	0.0664	0.1419**				
	(0.0676)	(0.0598)				
Ln(Share Workers, Services)			0.3227***	0.3147***		
			(0.0759)	(0.0600)		
Ln(Share Workers, Public Adm)					0.1608**	0.2487***
					(0.0619)	(0.0481)
% of Total Effect Mediated	0.007	0.071	0.252	0.247	0.107	0.402

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The outcome is log male literacy in the respective year. These cross-sectional models include province fixed effects; social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A15. Mediators: English Literacy, OLS

	(1)	(2)	(3)	(4)	(5)	(6)
	Ag Income		Land Taxes Per-Capita		Income Taxes Per-Capita	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0200***	0.0210***	0.0219***	0.0204***	0.0139***	0.0112***
	(0.0043)	(0.0035)	(0.0038)	(0.0032)	(0.0037)	(0.0029)
Ln(Ag. Income)	-0.0328	0.0606				
	(0.1302)	(0.1092)				
Ln(Land Taxes Per-Capita)			-0.1147	-0.0563		
			(0.1049)	(0.0994)		
Ln(Income Taxes Per-Capita)					0.4697***	0.5061***
					(0.0769)	(0.0597)
% of Total Effect Mediated	-0.0132	0.0128	-0.0221	-0.00582	0.363	0.435

	Share Workers					
	Industry		Services		Public Adm	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0210***	0.0205***	0.0173***	0.0169***	0.0194***	0.0152***
	(0.0040)	(0.0032)	(0.0038)	(0.0028)	(0.0040)	(0.0031)
Ln(Share Workers, Industry)	0.3366**	0.4427***				
	(0.1579)	(0.1342)				
Ln(Share Workers, Services)			0.8877***	1.0457***		
			(0.1756)	(0.1223)		
Ln(Share Workers, Public Adm)					0.4866***	0.6959***
					(0.1444)	(0.1049)
% of Total Effect Mediated	0.0098	0.0596	0.182	0.222	0.0853	0.303

Note: Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$ The outcome is log English literacy in the respective year. These cross-sectional models include province fixed effects; social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A16. Railroads and Mediators

	(1)	(2)	(3)	(4)	(5)	(6)
	Ag Income		Land Taxes Per-Capita		Income Taxes Per-Capita	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0080***	0.0045*	0.0040	0.0020	0.0170***	0.0172***
	(0.0028)	(0.0026)	(0.0029)	(0.0027)	(0.0036)	(0.0034)
Obs	163	157	188	188	190	187
	Share Workers					
	Industry		Services		Public Adm	
	1901	1911	1901	1911	1901	1911
Railroad Years	0.0006	0.0029	0.0044**	0.0046**	0.0037*	0.0095***
	(0.0021)	(0.0018)	(0.0017)	(0.0019)	(0.0022)	(0.0023)
Obs	187	187	187	187	187	187

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
The cross-sectional models include province fixed effects; social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A17. Other Mediators, Total Literacy

	(1)	(2)	(3)	(4)
	Share Migrants	Share Europeans	Indicator Protestant Mission	Share Workers, Non-Rail Service
	1901	1901	1911	1911
Railroad Years	0.0003	0.0000	0.0045***	0.0047**
	(0.0008)	(0.0000)	(0.0016)	(0.0019)
Obs	203	203	203	187
	Total Literacy			
Railroad Years	0.0078***	0.0075***	0.0067***	0.0052***
	(0.0019)	(0.0018)	(0.0017)	(0.0014)
Share Migrants	-0.1228			
	(0.1563)			
Share Europeans		20.0387***		
		(6.4384)		
Indicator Protestant Mission			0.0822	
			(0.0776)	
Ln(Share Workers, Non-Rail Service)				0.3792***
				(0.0626)
% of Total Effect Mediated	-0.00383	0.0424	0.0521	0.255

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The cross-sectional models include province fixed effects; social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.

Table A18. Other Mediators, Secondary Enrolment

	(1)	(2)	(3)	(4)
	Share Migrants	Share Europeans	Indicator Protestant Mission	Share Workers, Non-Rail Service
	1901	1901	1911	1911
Railroad Years	0.0003	0.0000	0.0045***	0.0047**
	(0.0008)	(0.0000)	(0.0016)	(0.0019)
Obs	203	203	203	187
		Secondary Enrolment		
Railroad Years	0.0156***	0.0161***	0.0139***	0.0128***
	(0.0031)	(0.0030)	(0.0032)	(0.0031)
Share Migrants	-1.1835***			
	(0.4042)			
Share Europeans		62.0239***		
		(12.5504)		
Indicator Protestant Mission			0.0096	
			(0.1564)	
Ln(Share Workers, Non-Rail Service)				0.4474***
				(0.1526)
% of Total Effect Mediated	0.0168	-0.0125	0.00134	0.123

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The cross-sectional models include province fixed effects; social controls namely the share of Brahmans, Christians, Muslims and Tribes; and GIS controls namely latitude, longitude, altitude, ruggedness, precipitation, distance to the coast, distance to a river and suitability for specific crops such as cotton, dryland rice, wetland rice, and wheat.